ENGINEERING-DRIVEN DATA ANALYTICS FOR IN SITU PROCESS MONITORING OF NANOMANUFACTURING

A Dissertation
Presented to
The Academic Faculty

by

Xiaowei Yue

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
[AUGUST, 2018]

COPYRIGHT © 2018 BY XIAOWEI YUE

ENGINEERING-DRIVEN DATA ANALYTICS FOR IN SITU PROCESS MONITORING OF NANOMANUFACTURING

Approved by:

Dr. Jianjun (Jan) Shi, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Ben Wang
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Chuck Zhang
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kamran Paynabar
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jionghua (Judy) Jin
Department of Industrial and Operations Engineering
The University of Michigan

Date Approved: [April 20, 2018]

To my beloved parents, wife Xing

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Jianjun (Jan) Shi, for his dedicated supervision, continuous presence and support, for his patience and insightful guidance throughout my Ph.D. studies. Without his support, I would have not been here. His words and deeds profoundly influenced my research. His kindness, patience, and enthusiasm have made a significant impact on my personality. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I would also like to thank my dissertation committee members, Professor Jionghua (Judy) Jin, Professor Kamran Paynabar, Professor Ben Wang, and Professor Chuck Zhang for their insightful comments and encouragement that improved my work from various perspectives. I would like to express my gratitude for their continuous support and for evaluating my dissertation.

I would like to thank my collaborators Dr. Jeffrey H. Hunt, Dr. Richard Liang, Dr. Jin Gyu Park, Dr. Rui Tuo, Dr. Kan Wang and Dr. Hao Yan, for their kind support and assistance and all the professors and friends who have shaped my academic life during my Ph.D. studies. Especially, I would like to express my appreciation to Ms. Jing Huang and Dr. Kaibo Wang who recommended me to join Professor Jianjun Shi's research group. I would also like to thank Dr. Yu Ding, Dr. Alan Erera, Dr. Nagi Gebraeel, Dr. Ran Jin, Dr. James Kong, Dr. Kaibo Liu, Dr. Matt Plumlee, Dr. Spiridon Reveliotis, Dr. Edwin Romeijn, Dr. Brani Vidakovic, Dr. Jeff Wu, Dr. Li Zeng for their kind support and assistance throughout my doctoral studies and job search.

I would like to express my gratitude to my fellow friends for the stimulating discussions and valuable suggestions, and for all the fun we have had in the last five years. They include, but are not limited to: Ms. Juan Du, Dr. Xiaolei Fang, Dr. Li Hao, Mr. Mohammad Nabhan, Dr. Matthew Plumlee, Mr. Mostafa Reisi, Ms. Xinran Shi, Dr. Andi Wang, Dr. Wenjia Wang, Dr. Yan Wang, Mr. Yuchen Wen, Dr. Weijun Xie, Dr. Hao Yan, Dr. Chen Zhang, Mr. Ruizhi Zhang, Dr. Tingyu Zhang, Mr. Zhen Zhong. I would like to express my gratitude and appreciation to Ms. Liping Luo for her kindness and tremendous care.

This journey would not have been possible without the love and support of my wife, Xing Wang. Last but not the least, I would like to thank my parents and my siblings for their spiritual support throughout my doctoral studies, and life in general. This dissertation stands as a testament to your unconditional love and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Carbon Nanotubes (CNTs) buckypaper is a multifunctional platform material with high potential for creating lightweight, high-performance materials for numerous applications due to its superior mechanical and electrical characteristics. One of the critical roadblocks to scale-up production of high-quality buckypaper is the monitoring and control of buckypaper made from the continuous nanomanufacturing process. There is a pressing need to address the fundamental research issues, develop transformative manufacturing technologies, and propose systematic methodologies to realize engineering-driven data analytics for in situ process monitoring of high-quality buckypaper.

On the basis of these initiatives, the dissertation focuses on developing systematic methodologies for effective data analytics, process monitoring and quality control in the continuous nanomanufacturing process. This dissertation started with an introduction to illustrate the motivation, research objectives, state-of-the-art and organization structure of the dissertation.

In Chapter 2, a novel generalized wavelet shrinkage method was proposed to realize real-time denoising and signal enhancement for nanomanufacturing process. Raman spectroscopy is an attractive in-line quality characterization and quantification tool because of its non-destructive nature, fast data acquisition speed, and ability to provide detailed material information. However, there is signal-dependent noise buried in the Raman spectra, which reduces the signal-to-noise ratio (S/N ratio) and affects the accuracy, efficiency, and sensitivity for Raman spectrum-based quality control approaches. Based on the validated signal-noise relationship, a generalized wavelet shrinkage approach was

introduced to remove noise in wavelet coefficients by applying individual adaptive wavelet thresholds. The effectiveness of this method was demonstrated using simulation and case study.

In Chapter 3, a penalized mixed-effects decomposition (PMD) was proposed to solve the multichannel profile detection problem in nanomanufacturing. The proposed PMD exploits a regularized high-dimensional regression with linear constraints to decompose the profiles into four parts: fixed effects, normal effects, defective effects, and signal-dependent noise. Finally, the separated fixed effects coefficients, the normal effects coefficients, and the defective effects coefficients can be used to monitor fabrication consistency, within-sample uniformity, and defect information, respectively. Using surrogated data analysis and case study, we evaluated the performance of the proposed method, and demonstrated a better detection power with less computational time.

In Chapter 4, we focused on quality inspection from massive high-dimensional Raman mapping data with mixed-effects. The existing tensor decomposition methods cannot separate mixed effects, and existing mixed-effects model can only handle vector/matrix data instead of multi-array data. We proposed a novel tensor mixed-effects (TME) model to analyze massive Raman mapping data with complex structure. The proposed TME model can (i) separate fixed effects and random effects in a tensor domain; (ii) explore the correlations along different dimensions; and (iii) realize efficient parameter estimation by a proposed double Flip-Flop algorithm. We also investigated the properties of the TME model, existence and identifiability of parameter estimation. The numerical analysis demonstrated the efficiency and accuracy of the parameter estimation. Convergence and asymptotic properties were discussed in the simulation and surrogate

data analysis. A real case study illustrated an application of the TME model in quantifying the influence of alignment on the CNTs buckypaper.

In summary, the dissertation contributes to the area of *System Informatics and Control* (SIAC) to develop systematic methodologies based on massive high-dimensional data for in situ process monitoring of nanomanufacturing. The fundamental methodologies developed in this thesis have potential to be applied to other advanced manufacturing systems.

CHAPTER 1. INTRODUCTION

1.1 Motivation

Carbon nanotubes (CNTs) buckypaper is a multifunctional platform material because it has great potential for creating lightweight, high-performance materials for various applications. To attain their intrinsic multifunctional properties at the macro-scale and scale-up production in continuous forms, CNTs must be effectively and efficiently aligned and covalently interconnected into the desired architecture on the nanoscale. An effective approach for achieving such goals is to make thin (5-25μm) CNTs sheets or "buckypaper" for use as reinforcement or functional layers that have high nanotube loading, good distribution/dispersion, high alignment, and desired inter-tube connections. The buckypaper has superior mechanical and electrical conductivity performance for applications such as conductive composites and lightweight electricity cable covers for EMI shielding. One of the critical roadblocks to scale-up production of high-quality buckypaper is the monitoring and quality control of buckypaper made from the continuous nanomanufacturing process. There is a pressing need to address these fundamental issues and propose systematic methodologies to realize in situ process monitoring and quality improvement for nanomanufacturing of high-quality buckypaper. Meanwhile, massive high-dimensional data with noise is generated during the nanomanufacturing process, how to do denoising, feature extraction and data analytics are very challenging and indispensable. On the basis of these initiatives, the dissertation focuses on developing systematic methodologies for engineering-driven data analytics, system modeling, process monitoring and quality improvement in the continuous nanomanufacturing process.

1.2 Research Objectives

The objectives of this research to develop systematic methodologies for data analytics and process monitoring for scalable nanomanufacturing of high-quality CNTs buckypaper. The proposed research focuses on three areas, which are:

(i) developing signal enhancement and denoising techniques for Raman spectra to realize in-line process monitoring and quality control of continuous nanomanufacturing;

(ii) constructing statistical models and extracting key features relevant to various quality characteristics of CNTs buckypaper;

(iii) establishing a tensor mixed-effects model for in-line monitoring based on massive high-dimensional Raman mapping data with complex correlation structure.

1.3 State-of-the-art

The most commonly used method to make buckypaper involves a filtration process in which a nanotube suspension in aqueous or organic media passes through a membrane filter and leaves a random quasi-two-dimensional network of carbon nanotubes or their bundles on the membrane. Continuous buckypaper manufacturing processes including pre-mixing, sonication, filtration, surfactant removal and drying have also been preliminarily explored. Numerous studies have been conducted on buckypaper fabrication and applications (Liang et al. 2008; Park et al. 2009; Wang et al. 2012). However, these studies mainly focused on the exploration of buckypaper synthesis and properties at the lab-scale.

Systematic methodologies of engineering-driven data analytics for in situ process monitoring were not founded in the literature.

In order to realize in-line monitoring of nanomanufacturing, an effective characterization method for nanostructure information should be explored. Raman spectroscopy has been widely used in the batch-to-batch nanomanufacturing of CNTs buckypaper materials (Gommans et al. 2000; Raravikar et al. 2002; Liu and Kumar 2003; Park et al. 2008). Recently, in-line Raman spectroscopy has been developed to monitor biochemical reactions and pharmaceutical crystallizations (Fevotte 2007; Alahbabi et al. 2006; Abu-Absi et al. 2011). In these efforts, in-line Raman spectra were found effective in providing detailed chemical molecular information. This non-destructive testing can be accomplished within seconds during continuous reactions or manufacturing processes. However, one of the major challenges of using in-line Raman spectroscopy in nanomanufacturing is the trade-off between data acquisition time and signal quality, i.e., signal-to-noise ratio (S/N ratio) due to the samples moving fast compared to static tests. A typical Raman characterization takes multiple scans with at least 10 seconds of acquisition time per scan. The procedure ensures a high S/N ratio, but takes 10 seconds to several minutes to acquire one spectrum (Beer et al. 2006; Ono et al. 2004). For in-line quality control of a nanomanufacturing process, a shorter data acquisition time with an adequate S/N ratio is more desirable. A tailored generalized wavelet shrinkage needs to be proposed to perform denoising and signal enhancement of in-line Raman spectra. The novel method can also be extended to other applications of removing noise with signal-dependent properties.

After we get the in-line Raman spectra, another key research topic is to model and analyze profile data for process monitoring and detection. Most of the researchers focused on linear profile modeling and monitoring (Kang and Albin 2000; Kim et al. 2003; Mahmoud et al. 2007). Nonlinear profile monitoring was also investigated by different techniques, such as polynomial regression, cubic spline, and B-spline (Mosesova et al. 2006; Zou et al. 2008; Jensen and Birch 2009; Shiau et al. 2009; Chang and Yadama 2010). However, the aforementioned approaches only focus on modeling of smooth and differentiable profiles, which are not applicable when profiles include spikes or peaks such as Raman spectra. From another point of view, researchers considered both fixed effects and random effects for profile monitoring (Mosesova et al. 2006; Jensen and Birch 2009; Shiau et al. 2009; Paynabar and Jin 2011). However, the mixed-effects model cannot be used directly in profile modeling and analysis of nanomanufacturing for following reasons: (i) A mixed-effects model can only measure between-profile variation by random effects, instead of separating normal between-profile variation and defective between-profile variation. The normal between-profile variation results from measurements and product uncertainty within the tolerance. While the defective between-profile variation is due to the defects in the buckypaper. (ii) There are too many coefficients used in a mixed-effects model, especially for high-dimensional profiles. This results in high false alarm rate and low detection rate in the multivariate process control. (iii) The algorithms to do parameter estimation of the mixed-effects model are not computationally efficient enough for in-line detection. Thus, we need to propose a novel penalized mixed-effects decomposition that overcomes these limitations.

Due to the recent development of metrology technologies, Raman mapping (also called Raman spectral imaging) has been used to perform in-line quality inspections in the continuous CNTs buckypaper fabrication processes. Raman mapping is a technique for generating detailed multi-array Raman spectra including numerous material information. Meanwhile, it is a challenging task to conduct data analytics, feature extraction and in-line decision making, due to the high-dimensionality, large data size, as well as complex spatial and temporal correlation structures of Raman mapping. According to the data structure, tensor is an efficient mathematical tool for formulating Raman mapping data. In high-dimensional statistics and machine learning, people used high-order tensor decompositions, such as CANDECOM/PARAFAC (CP) decomposition or Tucker decomposition to separate different components inherent to the data (Kolda and Bader 2009). Zhou et al. (2013) proposed a GLM model in the tensor domain, which extended the classical vector-valued covariate regression to an array-valued covariate regression. However, these tensor-based methods do not consider the multilevel variabilities (mixed effects) in the datasets. A mixed-effects model is a statistical model containing both fixed effects and random effects, and it is widely used in a variety of disciplines such as physics, biology, engineering and social sciences (Demidenko 2013; Galecki and Burzykowski 2013). However, the mixed effects model treats multivariate data as a vector or a matrix, which is insufficient for analysis of high-dimensional data, such as tensor-type Raman mapping, due to high dimensionality and complex correlations. Thus we need to develop a novel tensor mixed-effects (TME) model that can explore mixed effects in the tensor domain.

1.4 Organization of the Dissertation

This dissertation is organized in a multiple chapter format. Each of Chapters 2, 3, and 4 are written as a research paper, which has either been accepted or submitted for journal publications. Figure 1 looks out on the structure of this dissertation and also the relationship among these chapters.

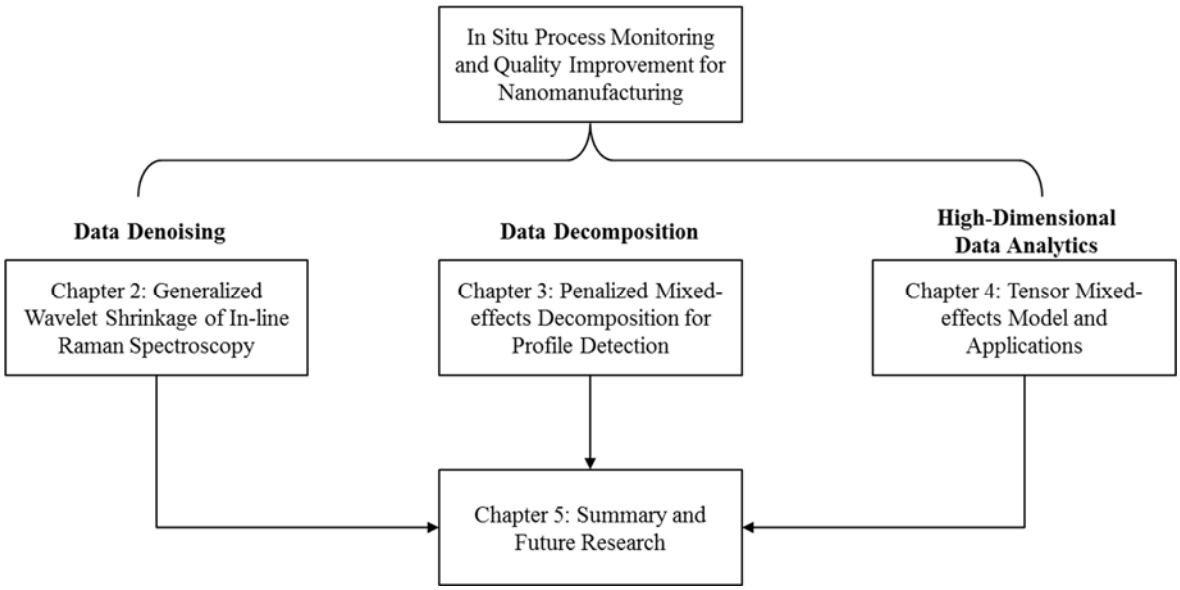


Figure 1. Outline of the Dissertation

In Chapter 2, we developed a signal analysis model with signal-dependent noise for Raman spectroscopy and validated the model based on experimental data. Based on the validated signal-noise relationship, a novel generalized wavelet shrinkage approach was introduced to remove noise in all wavelet coefficients by applying individual adaptive wavelet thresholds. The effectiveness of this method was demonstrated using both simulation and experimental case study of in-line Raman spectra. The proposed method allows for a significant reduction of Raman data acquisition time without much loss of S/N

ratio, which inherently enables Raman spectroscopy for in-line monitoring and control for continuous nanomanufacturing processes.

In Chapter 3, a novel penalized mixed-effects decomposition (PMD) was proposed to solve the multichannel profile detection problem in Raman spectroscopy. The proposed PMD exploited a regularized high-dimensional regression with linear constraints to decompose the profiles into four parts: fixed effects, normal effects, defective effects, and signal-dependent noise. An optimization algorithm based on the accelerated proximal gradient (APG) was developed to efficiently estimate parameters for the proposed model. Finally, the separated fixed effects coefficients, normal effects coefficients, and defective effects coefficients can be used to monitor fabrication consistency, within-sample uniformity, and defect information, respectively. Using the surrogated data analysis and the case study, we evaluated the performance of the proposed method, and demonstrated a better detection power with less computational time.

In Chapter 4, we proposed a tensor mixed effects (TME) model to analyze massive high-dimensional Raman mapping data with complex correlation structure. The proposed TME model can (i) separate fixed effects and random effects in a tensor domain; (ii) explore the correlations along different dimensions; and (iii) realize efficient parameter estimation by a proposed iterative double Flip-Flop algorithm. We also investigated the properties of the TME model, existence and identifiability of parameter estimation. The numerical analysis demonstrated the efficiency and accuracy of the parameter estimation in the TME model. Convergence and asymptotic properties were discussed in the simulation and surrogate data analysis. The real case study illustrated an application of the TME model in quantifying the influence of alignment on carbon nanotubes buckypaper.

In Chapter 5, we concluded the dissertation and summarized the original contributions. In addition, future research topics associated with the dissertation have also been discussed.

CHAPTER 2. GENERALIZED WAVELET SHRINKAGE OF IN-LINE RAMAN SPECTROSCOPY FOR QUALITY MONITORING OF CONTINUOUS NANOMANUFACTURING

Raman spectroscopy is an attractive in-line quality characterization and quantification tool for nanomanufacturing because of its non-destructive nature, fast data acquisition speed, and ability to provide detailed material information. However, there is signal-dependent noise buried in the Raman spectra, which reduces the signal-to-noise ratio (S/N ratio) and affects the accuracy, efficiency, and sensitivity for Raman spectrum-based quality control approaches. In this section, a signal analysis model with signal-dependent noise was developed and validated based on experimental data. The wavelet shrinkage method was used for denoising and improving the S/N ratio of raw Raman spectra. Based on the validated signal-noise relationship, a novel generalized wavelet shrinkage approach was introduced to remove noise in all wavelet coefficients by applying individual adaptive wavelet thresholds. The proposed method allows for a significant reduction of Raman data acquisition time without much loss of S/N ratio, which inherently enables Raman spectroscopy for in-line monitoring and control for continuous nanomanufacturing processes.

2.1 Introduction

As an effective off-line characterization method for nanostructure information, Raman spectroscopy has been widely used in batch-to-batch nanomanufacturing of CNTs buckypaper materials (Gommans et al. 2000; Raravikar et al. 2002; Liu and Kumar 2003;

Park et al. 2008). Recently, in-line Raman spectroscopy has been developed to monitor biochemical reactions and pharmaceutical crystallizations (Fevotte 2007; Alahbabi et al. 2006; Abu-Absi et al. 2011). In these efforts, in-line Raman spectra were found effective in providing detailed chemical molecular information. And the Raman spectra collection could be accomplished within seconds during continuous reactions or manufacturing processes. For similar purposes, we focus on developing an in-line Raman spectroscopy approach for quality monitoring and control of the continuous nanomanufacturing process.

One of the major quality control challenges of using in-line Raman spectroscopy is the trade-off between data acquisition time and signal quality, i.e., signal-to-noise ratio (S/N ratio). A typical Raman characterization ensures a high S/N ratio, but takes 10 seconds to several minutes to acquire one spectrum (Beer et al. 2006; Ono et al. 2004). For in-line quality control of the nanomanufacturing process, a shorter data acquisition time with an adequate S/N ratio is more desirable because of the following: (i) a shorter data acquisition time decreases the control bandwidth for process monitoring and control, which will correspondingly enhance the effectiveness and accuracy of quality monitoring and decrease losses from out-of-control situations; and (ii) it decreases the material heterogeneity dependent noise and provides a better in-line monitoring capability, as the materials to be inspected are moving quickly in a continuous buckypaper nanomanufacturing process.

A low S/N ratio in Raman spectra may cause several issues in feature extractions. If the magnitude of noise is comparable to the small peaks caused by impurities or defects in the buckypaper, important features could be overlooked. Even if the peaks are distinguishable, a small S/N ratio may cause large uncertainties in the determination of

peak positions, peak intensity and peak width, which results in the inaccuracy in the peak analysis. For example, in the Raman spectra for buckypaper, the intensity ratio of two peaks is the feature of interest, known as the intensity ratio of D-band and G-band. This ratio can be affected by the degree of functionalization. The intensity ratio of polarized Raman spectrum along the parallel and perpendicular directions to the alignment direction of carbon nanotubes (CNTs) in a buckypaper can be used to determine its degree of alignment (Cheng et al. 2009). In those cases, a low S/N ratio in the Raman spectra can be especially problematic since errors in the estimation of peak intensity may have a significant impact on the intensity ratio.

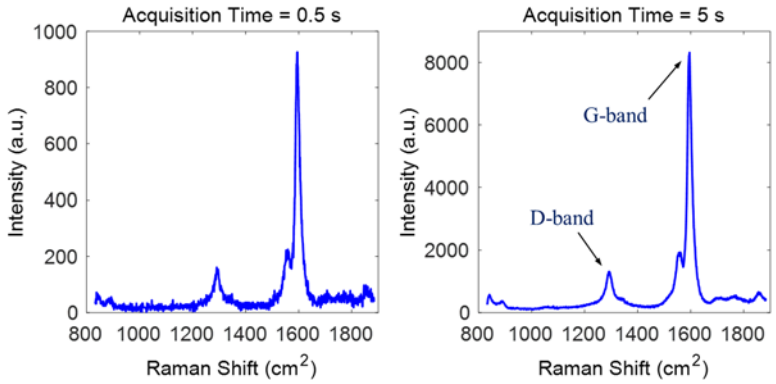


Figure 2. Typical Raman spectrum data of SWCNT buckypaper with a data acquisition time of (a) 0.5 seconds and (b) 5 seconds, respectively

Figure 2 shows the comparison between the Raman spectrum data taken from the same single-walled carbon nanotube (SWCNT) buckypaper sample with one pass of different acquisition time. The D-band and G-band are marked in Figure 2. Also, it is obvious that Raman spectrum with 5-second acquisition time has a higher S/N ratio than the one with 0.5-second acquisition time. This is because that Raman intensity tends to be large as the increase of the acquisition time, and the effect of random noise tends to be smaller. Detailed explanations can be found in Section 2.2.1 and 2.2.2.

If the characteristics of the noise are known, it is possible to design a specific filtering technique to take advantage of these characteristics and improve the S/N ratio through denoising. Typical denoising methods for nonlinear signals include spline and wavelet approaches. The spline basis is smooth, continuously differentiable and, therefore, has excellent approximation properties for smoothing signals (Wahba 1990). However, Raman spectra are not smooth in general and many peaks exist in the signal, as shown in Figure 2. Wavelet analysis is one of the most powerful and widely applied tools in the time-frequency transformation and denoising because of three advantages: (i) it is an effective non-parametric approach to characterize nonlinear profiles; (ii) it effectively extracts time-frequency features in multiple resolutions, which overcomes the limitations of conventional Fourier Transform (FT) and Short Time Fourier Transform (STFT) methods; and (iii) the obtained wavelet coefficients are sparse so that signals can be effectively denoised or compressed. Wavelet analysis has been widely used for spike removal, denoising, and signal compression for Raman spectra (Ramos and Ruisánchez 2005; Ehrentreich 2002). Conventional wavelet denoising methods (e.g., VisuShrink, RiskShrink, SureShrink) assume that noise in the signals is Gaussian white noise (Donoho and Johnstone 1995). A signal with correlated noise can be processed by a level-dependent wavelet shrinkage method (Johnstone and Silverman 1997).

A specific challenge in the buckypaper fabrication process is that Raman spectra are usually provided with signal-dependent noise. In this study, a generalized wavelet shrinkage method was proposed in order to make full use of the property of signal-dependent noise, and to denoise the entire Raman spectrum of CNTs buckypaper to enhance the signal of Raman spectra before feature extraction. Using this method, the data

acquisition time can be reduced without compromising the S/N ratio. This method enables in situ quality control based on Raman spectroscopy, and improves control accuracy, sensitivity, and efficiency. In addition, it can potentially improve the scalability of nanomanufacturing processes. To demonstrate the effectiveness of this method, we used a custom-made continuous Raman test setup and continuous buckypaper sample to conduct a case study. Both the simulation and the case study have revealed that the proposed approach outperforms conventional wavelet denoising methods and improves the S/N ratio by up to 100 percent, which means only less than half of the original data acquisition time is required to obtain signals with an equivalent S/N ratio.

The remainder of this article is organized as follows. Section 2.2 illustrates the noise source analysis for Raman spectrum, as well as the noisy signal modeling and validation; Section 2.3 describes a generalized wavelet shrinkage method and compares it with the conventional wavelet denoising method; Section 2.4 presents a simulation study to validate the proposed wavelet shrinkage method. Section 2.5 demonstrates how the proposed method can be used in in-line Raman characterization of buckypaper. Finally, a brief summary is discussed in Section 2.6.

2.2 Noise Pattern Analysis and Signal Modeling of Raman Spectrum

2.2.1 Noise Source Analysis

In order to develop denoising and signal enhancement techniques for Raman spectra of CNTs buckypaper materials, noise sources and patterns should be analyzed first. Table 1 summarizes five types of noise in Raman spectra, including photon shot noise,

sample-generated noise, instrument-generated noise, computationally generated noise, and externally generated noise (Pelletier 2003).

Table 1. Noise source and properties for different types of noise

Noise Types	Noise Sources	Characteristics
Photon shot noise	Statistical nature of light	Inescapable; Proportional to the intensity
Sample-generated noise	Sample composition ratio; Fluorescence	Temperature-dependent, heterogeneity-dependent(moving)
Instrument-generated noise	Detector noise such as thermal noise; variation in laser intensity	Temperature-dependent, provide useful information of the performance of the instrument
Computationally generated noise	Read out noise; converting electrons from the detector to a digital value	Independent with the signal magnitude and measurement time
Externally generated noise	Fluorescent room lighting, sunlight, cosmic rays	Independent with signal magnitude

Based on noise characteristics, we can propose a signal model of Raman spectra with the following facts or assumptions: (i) photon shot noise and sample-generated noise can be measured from noise in the Poisson process because the Raman scattering and photon arrival process follows a Poisson distribution and the influence of background counts is usually negligible (McCreery 2005); (ii) readout noise is the main element of instrument-generated noise and computationally generated noise, which is assumed to be additive Gaussian noise and is independent from the Poisson process; (iii) externally generated noise is negligible since specific measurements have been implemented to prevent effects of sunlight in the Raman instrument. This is verified because no spikes

caused by cosmic rays were found in the training spectra; and (iv) the temperature remains consistent during a short measurement time.

2.2.2 Signal Modeling of Raman Spectrum

A Raman instrument uses the CCD camera to identify the intensity of the scattered photons. The data collected by each frequency bin in a Raman instrument can be described by the Snyder's Model (Snyder et al. 1993):

$$y_i = n_i^s + g_i \tag{1}$$

where y_i denotes data collected by the i^{th} frequency bin in a Raman device; n_i^s is the number of photoelectrons generated on the i^{th} bin with ignoring the influence of background counts; g_i is the addictive Gaussian readout noise from the amplifier, and $n_i^s \sim poisson(u_i^s)$; $g_i \sim N(0, \varepsilon_g)$. n_i^s becomes larger as the acquisition time becomes longer. This can be observed in Figure 2. n_i^s and g_i are assumed statistically independent. For a Poisson distribution $poisson(\lambda)$ with sufficiently large values of λ (e.g. $\lambda > 1000$), the Poisson distribution can be approximated by a normal distribution. Meanwhile, $u_i^s > 10^7$ is satisfied in the Raman spectroscopy analysis (McCreery 2005). Thus, based on the normal approximation, $n_i^s \sim N(u_i^s, u_i^s)$, y_i follows a normal distribution with $N(u_i^s, u_i^s + \varepsilon_g)$.

The Raman intensity is correlated with data collected by each frequency bin in the Raman device. Thus, Equation (1) can be rewritten as Equation (2):

$$y(v) = s(v) + \sigma(s(v)) \cdot \varepsilon(v) \tag{2}$$

where $y(v)$ is the observed Raman intensity with respect to Raman shift v; $s(v)$ is the real Raman intensity; and $\sigma(s(v)) \cdot \varepsilon(v)$ is the signal-dependent noise with respect to Raman shift v in a Raman spectrum. $\sigma(\cdot)$ denotes the function of noise and Raman intensity; $\varepsilon(v)$ means the Gaussian white noise under Raman shift v.

If we consider a linear relationship between the Raman intensity and the collected data in a Raman detector (McCreery 2005), we can get that the $y(v)$ in Equation (2) follows a normal distribution $N(s(v), a \cdot s(v) + b)$. a and b can be obtained from historical data of the Raman spectra, which will be introduced in next sub-section.

The noise of Raman intensity under different Raman shifts is assumed to be uncorrelated, which is reasonable based on the validation test with real data shown in Section 2.2.3. We can get the signal model of a Raman spectrum in a matrix form:

$$\mathbf{Y} = \mathbf{S} + \mathbf{\Sigma} \sim N(\mathbf{S}, a \cdot \mathrm{diag}(\mathbf{S}) + b \cdot \mathbf{I}) \tag{3}$$

where **Y** is the observed Raman intensity vector, which is a column vector of $y(v)$ with respect to sequential Raman shifts; **S** is the real Raman intensity vector whose items are $s(v)$ with sequential Raman shifts; and **Σ** is the signal-dependent noise vector, that is a column vector of $\sigma(s(v)) \cdot \varepsilon(v)$ with respect to sequential Raman shifts.

2.2.3 Validation of Raman Spectroscopy Modeling

Before the application of the Raman spectrum model, a detailed validation is required for two purposes: (i) to verify the model effectiveness, including the validity of the assumption and constitution used in the modeling efforts; and (ii) to obtain the key parameters from the experimental data. In this validation effort, eight groups of experiments are designed and conducted with the parameters shown in Table 2. The experiments used two types of samples: single-walled carbon nanotubes (SWCNT) buckypaper and multi-walled carbon nanotubes (MWCNT) buckypaper. For the SWCNT buckypaper, each acquisition time setting (0.05 seconds, 0.1 seconds, 0.5 seconds, 1 second) is replicated 50 times in order to obtain a relatively accurate estimation of the variance in every point of the signal using a regular static Raman approach. Similarly, for the MWCNT buckypaper, different acquisition time settings (0.5 seconds, 1 second, 2 seconds, 5 seconds) are replicated for 50 times. Different acquisition time settings are used for the MWCNT and SWCNT buckypaper due to the different relationship between the S/N ratios and the acquisition time for the MWCNT and SWCNT. The details of the instrument and experiment processes are provided in Section 2.5.

To validate the uncorrelated assumption of the noise under different Raman shifts, we repeated measuring 100 Raman spectra for the same sample. The average spectrum was also computed. After that, the correlation matrix among the Raman intensities under different Raman shifts was calculated. We observed that the absolute value of the off-diagonal entries of the correlation matrix were consistently lower than 0.3. We then performed a hypothesis test with the null hypothesis that the noise was uncorrelated and the alternative hypothesis that the noise was correlated. The test result suggested that we could not reject the null hypothesis at a significance level of 0.05.

To test the signal-dependent property of noise, the sample variance for every Raman shift can be calculated. In Figure 3, we can see the curved shape of the calculated variance at each point is very similar to the tuning of signals by a linear transformation. By calculating the Pearson product-moment correlation coefficient (PPMCC) (Neter et al. 1996), the PPMCC of the variance vector and a SWCNT signal is 0.958 and the PPMCC of variance vector and a MWCNT signal is 0.901. Both of them are strongly correlated, which indicates that the noise is signal-dependent and the noisy signal model can fit the signal and variance well. Figure 3 shows the visual comparison between the variance and the transformed signal and regression relationship between Raman shift and variance for SWCNT with an acquisition time of 1 second. Moreover, based on the minimal mean square error principle, linear parameters a and b for every acquisition time setting of SWCNT and MWCNT have been validated. According to our validation results in Table 2, the value of linear parameters a and b remain stable for different materials and different acquisition time, which indicates that the linear parameters are only related with Raman instrument itself, but are independent of the choice of materials and acquisition time. The validated linear parameters for different materials and acquisition time settings remain that $a = 1; b = 41$.

In summary, the noise in Raman spectra, including photon shot noise, sample-generated noise, instrument-generated noise, computationally generated noise, and externally generated noise can be regarded as signal-dependent noise. And the dependence relationship remains stable and linear. The parameters are independent of different materials and acquisition time settings. Based on the training Raman spectra data from a buckypaper manufacturing process, the validated linear parameters are $a = 1; b = 41$.

Table 2. Design of experiment for the validation of Raman spectrum model

Sample Type	Acquisition Time	Validated Parameter
SWCNT	0.05 seconds	$a = 1; b = 42$
SWCNT	0.1 seconds	$a = 1; b = 41$
SWCNT	0.5 seconds	$a = 0.9; b = 44$
SWCNT	1 second	$a = 1; b = 41$
MWCNT	0.5 seconds	$a = 1; b = 41$
MWCNT	1 second	$a = 1; b = 41$
MWCNT	2 seconds	$a = 1; b = 41$
MWCNT	5 seconds	$a = 0.9; b = 42$

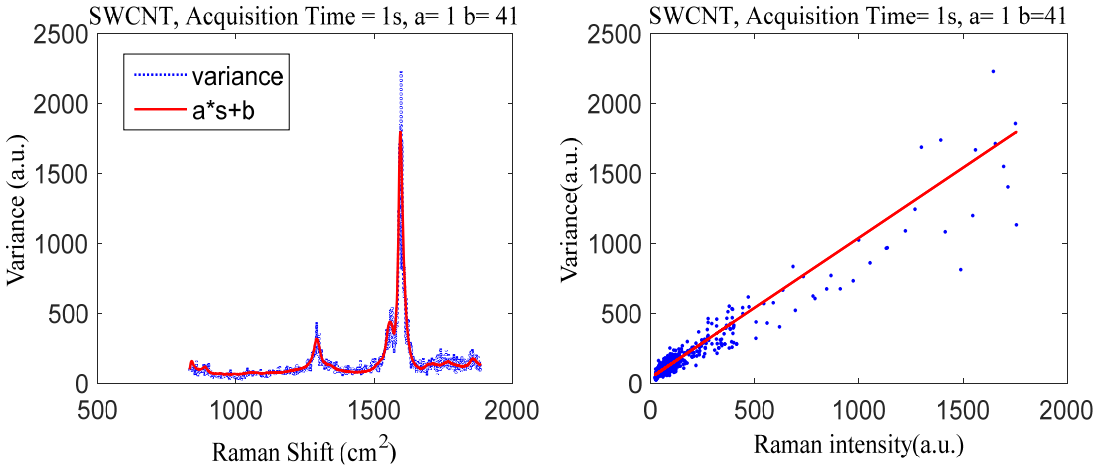


Figure 3. (a) Comparison between variance and transformed Raman spectrum data and (b) regression relationship between Raman shift and variance.

2.3 Generalized Wavelet Shrinkage Method

To reduce data acquisition time without loss of the S/N ratio or the improvement of the S/N ratio within a fixed data acquisition time, a generalized wavelet shrinkage method is developed in this study. The idea of the generalized wavelet shrinkage is to make full use of the validated linear dependence between the signal and the noise, and to promote the shrinkage from level-dependent to adaptive individual-dependent. That is to say,

conventional wavelet shrinkage methods determine thresholds based on different levels of wavelet coefficients. While for the generalized wavelet shrinkage, the threshold is individually determined based on each wavelet coefficient. We will first discuss the discrete wavelet decomposition, in Section 2.3.1, where Raman spectra are transformed into the wavelet domain. Then a conventional wavelet shrinkage method will be introduced in Section 2.3.2 to implement wavelet coefficients shrinkage based on level-dependent thresholds. With the help of the validated signal noise dependence, the signal variance can be estimated and an individual threshold can be determined for each wavelet coefficient. These "customized" thresholds will adaptively perform shrinkage corresponding to the wavelet coefficients and achieve a higher S/N ratio. The detailed technique of the generalized wavelet shrinkage method and its relationship with the penalized optimization will be introduced in Section 2.3.3. Section 2.3.4 implements the sensitivity analysis of the diagonal approximation. Finally, Section 2.3.5 provides further discussions on the criteria for selection of the penalized parameter and the wavelet basis.

2.3.1 Discrete Wavelet Decomposition

The wavelet transformation is selected to decompose Raman spectra because of the following characteristics: (i) wavelet-based modeling can handle complicated nonlinear Raman spectra; (ii) the wavelet-based method can analyze signals by multi-resolution, which provides the opportunity for the diverse resolution analysis on different segments of Raman spectra; and (iii) the scarcity of wavelet coefficients is suitable for soft shrinkage and hard shrinkage, which are better than linear shrinkage, especially for non-stationary and nonlinear signals.

The first step should be transforming every Raman spectra into a wavelet domain. According to the wavelet theory (Daubechies 1992), function $\phi(\cdot)$ and $\psi(\cdot)$are known as the scaling function (father wavelet) and wavelet function (mother wavelet) respectively. Usually, an orthonormal wavelet basis with basic properties is chosen. A standardized Raman spectrum data $g(v)$ is composed into two parts corresponding to coarse and detail parts. $g(v)$ belongs to the square-integral functions space $L^2(\mathcal{R})$ (Daubechies 1992).

$$g(v) = \sum_{k \in \mathbf{Z}} c_{sk}\phi_{sk}(v) + \sum_{j=s}^{\infty}\sum_{k \in \mathbf{Z}} d_{jk}\psi_{jk}(v)$$

where $\phi_{sk}(v)$ and $\psi_{jk}(v)$ are known as the scaling function and the wavelet function. c_{sk} and d_{jk} are called the approximate coefficients and the detail coefficients. They are determined by the inner product of $g(v)$ and a scaling function or a wavelet function.

According to the discrete wavelet transform algorithm, $g(v)$ can be divided into matrix product of wavelet coefficients and orthogonal wavelet basis. The matrix form of wavelet decomposition for the noisy signal model Equation (3) is

$$\mathbf{\Gamma} \cdot \mathbf{Y} = \mathbf{\Gamma} \cdot \mathbf{S} + \mathbf{\Gamma} \cdot \mathbf{\Sigma} \sim N(\mathbf{\Gamma} \cdot \mathbf{S}, a \cdot \mathbf{\Gamma} \cdot \mathrm{diag}(\mathbf{S}) \cdot \mathbf{\Gamma}^T + b \cdot \mathbf{I}) \quad (4)$$

where $\mathbf{\Gamma}$ is the wavelet transform matrix that depends on the selected wavelet basis.

Equation (4) can be rewritten as

$$\mathbf{W} = \mathbf{G} + \mathbf{E} \sim N(\mathbf{G}, \mathbf{\Lambda}) \quad (5)$$

where $\mathbf{W} = \mathbf{\Gamma} \cdot \mathbf{Y}, \mathbf{G} = \mathbf{\Gamma} \cdot \mathbf{S}, \mathbf{\Lambda} = a \cdot \mathbf{\Gamma} \cdot \mathrm{diag}(\mathbf{S}) \cdot \mathbf{\Gamma}^T + b \cdot \mathbf{I}$.

The scalar form of Equation (5) is

$$w_i^J = g_i^J + e_i^J, i = 1, \dots, c^J$$

where w_i^J denotes the i^{th} coefficients in the wavelet decomposition level J transformed from observed signals; g_i^J denotes the i^{th} coefficients ingredients with respect to real signal in the wavelet decomposition level J. e_i^J means the i^{th} coefficients ingredients with respect to signal-dependent noise in the wavelet decomposition level J.

2.3.2 Wavelet Shrinkage Method

One of the most popular wavelet-based analysis methods is wavelet shrinkage for signal denoising and compression. This wavelet shrinkage method is asymptotically optimal for recovering objects taken from certain functional classes, such as the Besov spaces (Candes 2006). As for multivariate normal distribution, hard shrinkage and soft shrinkage are two widely used thresholding rules, which correspond to the minimum complexity estimates for complexity penalties.

In order to fully understand the process, we firstly consider ℓ_2 norm penalty,

$$\min_{\mathbf{W}}\{(\mathbf{Y} - \mathbf{\Gamma}^T\mathbf{W})^T(\mathbf{Y} - \mathbf{\Gamma}^T\mathbf{W}) + \lambda\mathbf{W}^T\mathbf{W}\} \quad (6)$$

where $\mathbf{\Gamma}$ is a wavelet transform matrix that depends on the selected wavelet basis, $\mathbf{Y}$ denotes the original signals matrix, whose column means one Raman spectrum. $\mathbf{W}$ denotes wavelet coefficients matrix in the wavelet domain. λ is a penalty parameter.

Because of the orthonormality of wavelet basis matrix, $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}$, we can obtain the solution of Equation (6) as

$$\widehat{w}_i^J = \frac{1}{1+\lambda} w_i^J \tag{7}$$

where w_i^J denotes original the i^{th} coefficients in the wavelet decomposition level J transformed from the observed signals, $\widehat{w}_i^J$ denotes the estimated (after-denoised) i^{th} coefficients in the wavelet decomposition level J.

Straightforwardly, Equation (6) has a quadratic penalty and corresponds to the linear shrinkage. Linear shrinkage cannot remove noise effectively in wavelet coefficients, so it can be difficult to find its usage in wavelet-based methods in the literature.

The soft shrinkage method is closely correlated with the minimum complexity estimates with an ℓ_1 norm penalty, which is similar to LASSO (Candes 2006).

$$\min_{\mathbf{W}}\left\{(\mathbf{Y}-\mathbf{\Gamma}^T\mathbf{W})^T(\mathbf{Y}-\mathbf{\Gamma}^T\mathbf{W}) + \lambda\|\mathbf{W}\|_{\ell_1}\right\} \tag{8}$$

The soft shrinkage method can solve Equation (8) with

$$\widehat{w}_i^J = \eta_s\left(w_i^J, \frac{\lambda}{2}\right) = \text{sign}\left(w_i^J\right)\left(\left|w_i^J\right| - \frac{\lambda}{2}\right)_+ \tag{9}$$

where $\eta_s(\cdot)$ denotes the soft shrinkage function, $\text{sign}(\cdot)$ means the sign function, $\left(\left|w_i^J\right| - \lambda/2\right)_+ = \left(\left|w_i^J\right| - \lambda/2\right) \cdot I\left(\left|w_i^J\right| - \lambda/2\right)$ and $I(\cdot)$ is the indicator function. $\lambda/2$ is the threshold in the soft shrinkage function.

Similarly, the hard shrinkage

$$\widehat{w}_i^J = \eta_h\left(w_i^J, \sqrt{\lambda}\right) = \text{sign}\left(w_i^J\right) \cdot \left(\left|w_i^J\right|\right) \cdot I\left(\left|w_i^J\right| - \sqrt{\lambda}\right) \tag{10}$$

is the solution to the minimum complexity estimates with an ℓ_0 norm penalty and here $\sqrt{\lambda}$ is the threshold (Candes 2006).

$$\min_{\mathbf{W}}\left\{(\mathbf{Y} - \mathbf{\Gamma}^T\mathbf{W})^T(\mathbf{Y} - \mathbf{\Gamma}^T\mathbf{W}) + \lambda\|\mathbf{W}\|_{\ell_0}\right\} \tag{11}$$

One explanation for the fact that a soft shrinkage function is preferred for smoothing is that it possesses various statistical advantages like continuity of the rule. While the hard shrinkage rule yields keep-or-kill estimation, it will protect larger values in original signals, especially for peak values. In a buckypaper manufacturing process, D-band intensity peak and G-band intensity peak in a Raman spectrum is of great interest for monitoring and control. Therefore, the hard shrinkage method is preferred in our application.

The selection of threshold is one of the most important factors in the efficiency of wavelet denoising. Donoho and Johnstone proposed three popular thresholding methods: VisuShrink, RiskShrink, SureShrink (Donoho and Johnstone 1995; Johnstone and Silverman 1997) and compared them with the minimax paradigm. The major difference among these three methods is the procedure to get the threshold. VisuShrink applies the universal threshold derived under the constraint that with a high probability, the estimate should be at least as smooth as the signal; RiskShrink minimizes the theoretical upper

bound of the asymptotic risk; and SureShrink minimizes the Stein unbiased estimate of risk for threshold estimates.

However, a common limitation of these conventional wavelet shrinkage methods is that the threshold can only be determined globally or level-dependently. In other words, those methods use only the same threshold for all the wavelet coefficients in a given level. The variance of wavelet coefficients is estimated from the data directly. Thus, the intrinsic properties determine that the conventional wavelet shrinkage method cannot determine the individual threshold for each wavelet coefficient. Instead, according to the discussion in Section 2.2 validation, we know that the noise of Raman spectra in buckypaper manufacturing processes has signal-dependent characteristics. Furthermore, the relationship of the signal-noise dependence remains stable and linear. Thus, we can estimate the noise based on a linear parameter from the validation process, which enables us to create a corresponding threshold for each wavelet coefficient. We will develop a novel wavelet shrinkage method to extend level-dependent thresholding to individual-dependent thresholding.

2.3.3 Generalized Wavelet Shrinkage Method

A generalized wavelet shrinkage method is proposed to consider effects from the covariance matrix $a \cdot \mathbf{\Gamma} \cdot \mathrm{diag}(\mathbf{S}) \cdot \mathbf{\Gamma}^T + b \cdot \mathbf{I}$ of the multivariate Gaussian distribution in the wavelet-based analysis. However, there are two major differences between the generalized smoothing spline and the generalized wavelet shrinkage method: (i) The transformation basis matrices are different. One uses a wavelet basis, and the other a smoothing spline basis. In general, the coefficients in the wavelet domain are sparse, while

coefficients in the spline domain are not sparse. Thus, a smoothing spline is suitable for fitting stationary smoothing signals, while wavelet analysis is more powerful for non-stationary, nonlinear noisy signals like Raman spectra; and (ii) A generalized smoothing spline penalizes ℓ_2 norm, while a generalized wavelet shrinkage penalizes ℓ_1 norm or ℓ_0 norm. The reason of the different penalized order has been introduced in the previous subsection.

Considering the effect of covariance matrix $\mathbf{\Omega} = a \cdot \mathbf{\Gamma} \cdot \mathrm{diag}(\mathbf{S}) \cdot \mathbf{\Gamma}^T + b \cdot \mathbf{I}$, the wavelet-based minimum weighted sum of square estimate with a penalty ℓ_1 norm is

$$\min_{\mathbf{W}} \left\{ (\mathbf{Y} - \mathbf{\Gamma}^T \mathbf{W})^T \mathbf{\Omega}^{-1} (\mathbf{Y} - \mathbf{\Gamma}^T \mathbf{W}) + \lambda \|\mathbf{W}\|_{\ell_1} \right\} \tag{12}$$

Based on section 2.3.2, we know, it should correspond to the soft shrinkage. The solution of Equation (12) can be obtained as

$$\min_{\mathbf{W}} \left\{ (\mathbf{Y} - \mathbf{\Gamma}^T \mathbf{W})^T \mathbf{\Gamma}^T \mathbf{\Gamma} \mathbf{\Omega}^{-1} \mathbf{\Gamma}^T \mathbf{\Gamma} (\mathbf{Y} - \mathbf{\Gamma}^T \mathbf{W}) + \lambda \|\mathbf{W}\|_{\ell_1} \right\}$$

$$\min_{\mathbf{W}} \left\{ (\mathbf{\Gamma Y} - \mathbf{W})^T \mathbf{\Gamma} \mathbf{\Omega}^{-1} \mathbf{\Gamma}^T (\mathbf{\Gamma Y} - \mathbf{W}) + \lambda \|\mathbf{W}\|_{\ell_1} \right\}$$

where $\mathbf{\Omega}$ is a diagonal positive definite matrix, and $\mathbf{\Gamma}$ is an orthonormal wavelet basis set. Let $\mathbf{K} = \mathbf{\Gamma} \mathbf{\Omega}^{-1} \mathbf{\Gamma}^T$, since weak non-diagonal entries (most all of them are less than 10^{-4}, and the maximum is 10^{-3}) are more than 15 times less than strong diagonal entries (the mean of diagonal entries is 0.015), $\mathbf{K}$ can be approximated with a diagonal matrix with an acceptable accuracy. In Section 2.3.4, we will implement the sensitivity analysis of errors from the approximation by ignoring off-diagonal entries and describe the feasibility of the approximation.

The scalar solution of Equation (12) is a generalized soft shrinkage:

$$\widehat{w}_i^J = \eta_s\left(w_i^J, \frac{\lambda}{2k_i^J}\right) = \text{sign}\left(w_i^J\right)\left(\left|w_i^J\right| - \frac{\lambda}{2k_i^J}\right)_+ \tag{13}$$

where k_i^J is a corresponding diagonal of matrix $\mathbf{K} = \mathbf{\Gamma\Omega}^{-1}\mathbf{\Gamma}^T$, and $\lambda/2k_i^J$ is the threshold of the soft shrinkage. $\lambda/2k_i^J$ is proportional to the standard deviation of w_i^J.

With the penalty of ℓ_0 norm, the minimal weighted sum of square estimate should solve the optimization of Equation (14)

$$\min_{\mathbf{W}}\left\{(\mathbf{Y} - \mathbf{\Gamma}^T\mathbf{W})^T\mathbf{\Omega}^{-1}(\mathbf{Y} - \mathbf{\Gamma}^T\mathbf{W}) + \lambda\|\mathbf{W}\|_{\ell_0}\right\} \tag{14}$$

The generalized hard shrinkage is the solution:

$$\widehat{w}_i^J = \eta_h\left(w_i^J, \frac{\sqrt{\lambda}}{k_i^J}\right) = \text{sign}\left(w_i^J\right)\cdot\left(\left|w_i^J\right|\right)\cdot I\left(\left|w_i^J\right| - \frac{\sqrt{\lambda}}{k_i^J}\right) \tag{15}$$

Remark:

(1) The main idea of the generalized wavelet shrinkage method is to determine thresholds for shrinkage adaptively. Conventional wavelet shrinkage applies level-dependent thresholds to shrinkage wavelet coefficients since the variance cannot be estimated in each coefficient. The level-dependent threshold is suitable for a signal model with white noise. However, for signal-dependent noise in Raman spectra, the level-dependent threshold is not optimal because wavelet coefficients within the same level share the same threshold, although they correspond to different noise amounts. Due to the stable

and linear relationship between signal and noise variance, we can estimate the variance for each coefficient based on the validated signal model of Raman spectra. Then a generalized wavelet shrinkage method is designed to adjust the threshold individually for each coefficient, which outperforms the level-dependent threshold. It is notable that individual-dependent wavelet thresholds in the generalized wavelet shrinkage do not mean that the thresholding parameters can be freely changing and individually adjusted. Different shrinkage thresholds for different wavelet coefficients are mainly caused by the noise level through the weighting matrix $\mathbf{\Gamma\Omega^{-1}\Gamma}^T$, and the shrinkage parameter λ, which is fixed through a Phase-I analysis.

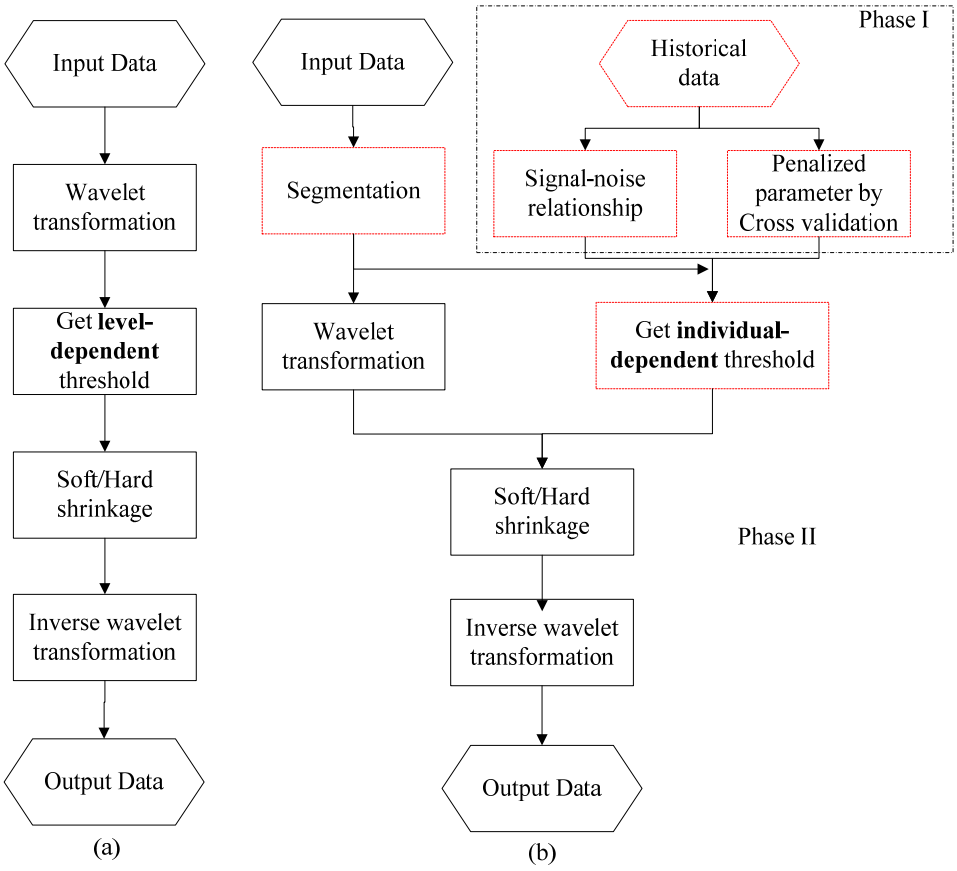
(2) The procedure for conventional wavelet shrinkage methods is shown in Figure 4(a). The Raman spectra signals (input signals) are transformed into a wavelet domain first, and then level-dependent thresholds are estimated based on wavelet coefficients. Soft or hard shrinkage methods then adjust the values of wavelet coefficients, and the adjusted wavelet coefficients are inversely transformed into denoised Raman spectra signals (output signals). The procedure for the generalized wavelet shrinkage method can be separated into two phases, as shown in Figure 4(b): (i) during phase I, historical signals can be used to validate the signal model of Raman spectrum and get the parameters that indicate variance-signals relationship. Besides, a penalized parameter can be determined based on cross validation in phase I; (ii) for phase II, input signals can be divided into several segmentations according to different feature locations. For example, the D-band intensity peak and the G-band intensity peak are of interest in a continuous buckypaper manufacturing process, so these two segments can be specified. After the signal segmentation, a wavelet transformation is implemented to get wavelet coefficients. An

individual wavelet threshold can be determined by the signal-noise relationship and penalized parameter from phase I and signals from phase II. The main difference between the conventional wavelet shrinkage method and the generalized wavelet shrinkage method are whether individual-dependent thresholding is implemented or not, which has been highlighted by the red dotted frames.

(3) Assuming the noisy signals have no signal-dependent noise, the covariance matrix $\mathbf{\Omega}$ will be a scalar matrix (a scalar matrix is a diagonal matrix that has equal main diagonal entries). Then the thresholds for different wavelet coefficients $\lambda/2k_i^J$ will be the same. In this case, the generalized wavelet shrinkage will be equivalent to the wavelet shrinkage. Also, the stronger the signal-noise dependence property is, the better performance the generalized wavelet shrinkage has comparing to the conventional wavelet shrinkage.

(4) The generalized wavelet shrinkage method can also be used to build the calibration model or preprocess of the calibration spectroscopic samples. Usually, multivariate statistic techniques, such as principal component regression, partial least squares regression, or Gaussian process regression are used to build the calibration model (Chen et al. 2007, Clegg et al. 2009). However, these methods are with the assumption that the spectroscopic data has Gaussian white noise. If wavelet coefficients are used as corresponding explanatory variables in a calibration model, the generalized wavelet shrinkage approach can be used to build a calibration model for spectroscopic data with signal-dependent noise. Besides, if the corresponding explanatory variables in calibration are not wavelet coefficients, the generalized wavelet shrinkage can be applied to preprocess

calibration spectroscopic samples with signal-dependent noise, which is able to prevent over-fitting of calibration modeling and increase the model accuracy.



(Note: The main difference between a conventional wavelet shrinkage method and the generalized wavelet shrinkage method are highlighted by red dotted frame.)

Figure 4. Procedure for (a) the conventional wavelet shrinkage and (b) the generalized wavelet shrinkage.

2.3.4 Sensitivity Analysis of Diagonal Approximation

In this section, we will implement the sensitivity analysis of errors from the approximation by ignoring off-diagonal entries, and then describe the feasibility of the approximation. Inspired by the diagonally dominant matrix, in which $|K_{ii}| > \sum_{j\neq i} |K_{ij}|$ for all i, we define $\mathrm{Q} = \frac{1}{p}\sum_{i=1}^{p}\left(|K_{ii}|/\sum_{j\neq i}|K_{ij}|\right)$ to quantify the relative value of the diagonal

entries compared to the summation of off-diagonal row entries. Q can also be regarded as the degree of diagonal dominance. For example, it is straightforward to show that for a diagonal dominant matrix, $Q \geq 1$. To perform the sensitivity analysis due to the approximation by ignoring off-diagonal entries under different Q. Similar to the simulation study, the spectrum with signal-dependent noise is used for this sensitivity analysis. The generalized wavelet shrinkage with the diagonal approximation is used to denoise the original signal and obtain the estimated profile $\hat{y}_1$. Since the soft thresholding corresponds to the convex ℓ_1 norm penalty, we can also solve the original problem without diagonal approximation by applying convex optimization directly and obtain the estimated profile $\hat{y}_2$. We define the relative difference between $\hat{y}_1$ and $\hat{y}_2$ as $\text{DIFF} = \sum_{i=1}^{p}(|\hat{y}_{1i} - \hat{y}_{2i}|) / \sum_{i=1}^{p}(|\varepsilon_i|)$. The sensitivity relationship between DIFF and Q is shown in Figure 5. From Figure 5, we can conclude that the DIFF becomes smaller as the Q becomes larger, which is consistent with our intuition that the influence of diagonal approximation will be small if the diagonal entries are much larger than the off-diagonal entries. If the Q is smaller than 30, the DIFF quickly increases as the decrease of Q. In our application on Raman spectra, the Q is 77.61 (marked as a red star in Figure 5.), which results in about 10.6% difference of the accuracy corresponding to the noise standard deviation. The difference is about 1.02 unit in Raman intensity, which is within the monitoring resolution based on domain knowledge.

Though the diagonal assumption leads to some approximation errors (less than 1.02 unit in this case), the proposed algorithm has significant computational advantages. The computation time of denoising one spectrum by our generalized wavelet shrinkage (around 0.0009 seconds) is much faster than the computational time of using the convex

optimization (around 1.0280 seconds). During the continuous manufacturing of the Carbon Nanotube Buckypaper, multiple profiles will be measured and the computation time of using the convex optimization cannot satisfy the in-line monitoring requirement, while the proposed generalized wavelet shrinkage can process the profiles efficiently and provide very similar denoising results.

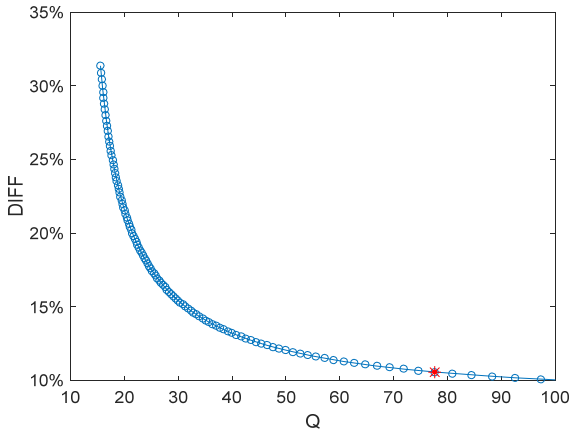


Figure 5. Sensitivity analysis between DIFF and Q

2.3.5 Criteria for Selection of the Penalized Parameter and the Wavelet Basis

2.3.5.1 Determine the penalized parameter by Revised Cross-Validation

This section describes how to obtain the penalized parameter in Phase I by using in-control, historical Raman spectra signals. Similar to the leave-one-out cross-validation (Neter et al. 1996), the proposed approach balances the trade-off between the bias and the variance of estimated parameters. The vital difference between the proposed approach and the regular leave-one-out cross-validation is the choice of the benchmark vector in the optimization equation.

The proposed optimization equation is given as

$$\mathbf{W}_{n,\lambda}^{[k]} = \underset{\mathbf{W}^{[k]}}{\operatorname{argmin}} \left\{ \left(\mathbf{Y}^{[k]} - \mathbf{\Gamma}^T \mathbf{W}^{[k]} \right)^T \mathbf{\Omega}^{-1} \left(\mathbf{Y}^{[k]} - \mathbf{\Gamma}^T \mathbf{W}^{[k]} \right) + \lambda \left\| \mathbf{W}^{[k]} \right\|_{\ell_0} \right\} \quad (16)$$

where $k = 1,2, \ldots, n$; $\mathbf{Y}^{[k]}$ denotes original signals matrix except for the k^{th} signal. $\mathbf{W}^{[k]}$ means corresponding wavelet coefficients matrix without wavelet coefficients for the k^{th} signal. $\mathbf{W}_{n,\lambda}^{[k]}$ represents the optimal wavelet coefficients matrix under signal number n and penalized parameter λ. The optimal penalized parameter λ suffices the minimum mean sum of square of the difference between estimated partial signals and ideal signals.

$$\lambda^* = \underset{\lambda \in \mathbf{R}^+}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(\overline{\mathbf{Y}} - \mathbf{\Gamma}^T \mathbf{W}_{n,\lambda}^{[k]}(i) \right)^T \left(\overline{\mathbf{Y}} - \mathbf{\Gamma}^T \mathbf{W}_{n,\lambda}^{[k]}(i) \right) \quad (17)$$

where λ^* denotes the optimal penalized parameter. $\overline{\mathbf{Y}}$ represents the mean vector of all the Raman spectra by an ensemble average. Ensemble average increases the S/N ratio based on the historical multiple measurements, so $\overline{\mathbf{Y}}$ can be regarded as the estimated ideal signals. $\mathbf{W}_{n,\lambda}^{[k]}(i)$ indicates the i^{th} column of $\mathbf{W}_{n,\lambda}^{[k]}$, that represents wavelet coefficients vector of the i^{th} signal.

The vital difference between the proposed approach and the regular leave-one-out cross-validation is that applying $\overline{\mathbf{Y}}$ as an important benchmark vector in the optimization equation, instead of the k^{th} signal in the regular leave-one-out cross validation. The reason for doing this is that an ensemble average $\overline{\mathbf{Y}}$ with high S/N ratio can be used to estimate ideal signals. Our target is picking up the optimal penalized parameter that enables the optimal denoising and signal enhancement instead of signal prediction.

Table 3. Performance comparison among different wavelet basis

Wavelet basis		λ	Mean $\lvert P_1 - P_1^* \rvert$	Std $\lvert P_1 - P_1^* \rvert$	Mean $\lvert P_2 - P_2^* \rvert$	Std $\lvert P_2 - P_2^* \rvert$
Raw Data	/	/	42.18	13.98	88.14	47.97
Coi	1	0.32	9.77	5.64	25.60	14.27
Coi	2	0.52	8.38	5.32	28.40	23.92
Coi	3	0.52	6.98	5.26	44.92	22.20
Coi	4	0.52	9.33	5.62	39.51	19.70
Coi	5	0.52	6.48	4.16	24.49	18.84
DB	4	0.32	10.31	7.15	37.94	34.97
DB	6	0.52	5.79	3.90	60.60	37.58
DB	8	0.36	7.88	6.35	41.94	27.39
DB	10	0.32	11.38	7.83	30.12	22.34
DB	12	0.52	16.35	5.20	31.50	21.25
DB	14	0.38	14.71	6.70	41.03	26.28
DB	16	0.34	12.42	7.28	24.34	20.59
DB	18	0.48	5.79	3.95	21.21	15.88
DB	20	0.50	7.58	5.43	59.69	19.56
Sym	4	0.52	12.90	5.46	34.04	26.66
Sym	5	0.36	9.49	8.12	27.17	25.99
Sym	6	0.52	12.78	5.55	33.50	24.58
Sym	7	0.34	8.93	8.02	25.92	20.32
Sym	8	0.52	14.93	5.49	25.90	20.71
Sym	9	0.50	6.40	4.08	29.33	27.72
Sym	10	0.52	6.00	4.78	24.71	16.26

* "DB" means "Daubechies wavelet", "Coi" denotes"Coiflet wavelet", 'Sym' denotes "Symmlet wavelet".

2.3.5.2 Determine the wavelet basis

In practical applications, several factors should be considered when choosing the wavelet basis. They are orthogonality, complexity, width, and shape. Orthogonal wavelets produce wavelet spectra that contain discrete "blocks" of wavelet power and give the most compact representation of the signal. Complex wavelet functions return information about both amplitude and phase. Width will affect the resolution of wavelet. In addition, the wavelet basis should reflect the type of features occurred in the signal (Torrence and Compo 1998). According to the four factors above and the curve shape of Raman spectra, Coiflet, Daubechies and Symmlet families are chosen and compared to analyze Raman spectra.

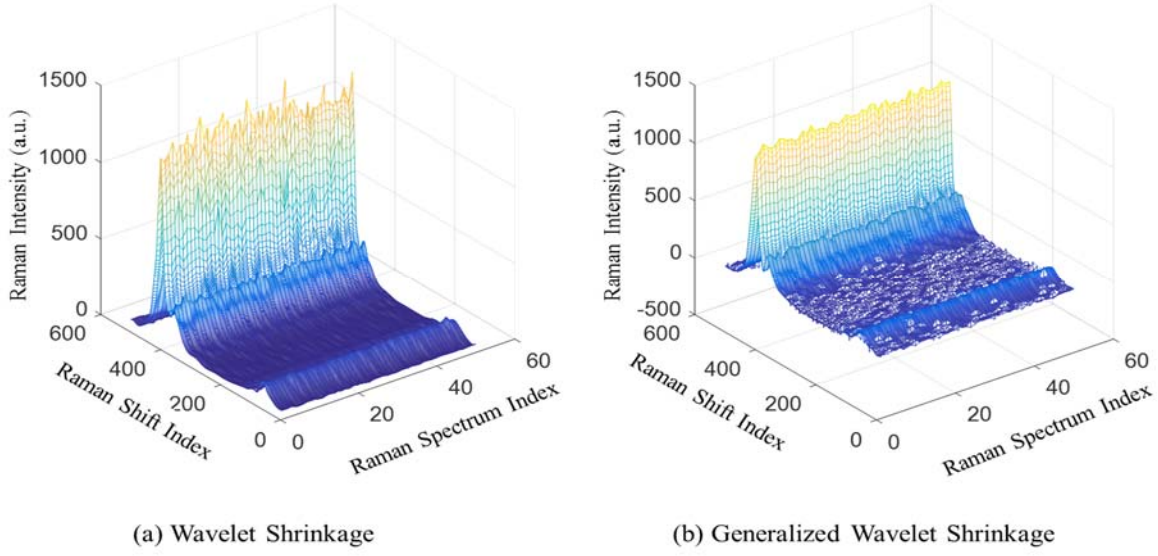
Further wavelet basis selection is introduced as following. As for each wavelet basis, the optimal penalized parameter is obtained by using the cross validation method. The target of wavelet denoising and signal enhancement is to get the accurate D-band peak intensity and G-band peak intensity. Thus, the difference between each estimated the optimal peak intensity (from ensemble average) and estimated denoising peak intensity are calculated during phase I, that is shown in Table 3.

In Table 3, P_1^* denotes the estimated the optimal D-band peak intensity by ensemble average; P_1 is the estimated D-band peak intensity from denoised signals by using the optimal penalized parameter λ^*; P_2^* and P_2 are corresponding notation for the G-band peak intensity. Four performance measures including mean of $|P_1 - P_1^*|$, standard deviation of $|P_1 - P_1^*|$, mean of $|P_2 - P_2^*|$, standard deviation of $|P_2 - P_2^*|$ are considered totally to determine the optimal wavelet basis as Daubechies 18.

2.4 Simulation

We use the average of smoothing Raman spectra as the ideal signal. The D-band peak intensity value $P_1^* = 151.3$, and the G-band peak intensity value $P_2^* = 1231$. There are 50 Raman spectra generated by adding signal-dependent noise. In the simulation, the conventional wavelet denoising is implemented using procedures in Figure 4(a) while the generalized wavelet shrinkage approach adopts procedures in Figure 4(b). Each signal is picked up as phase II signal in turn and the remaining 49 Raman spectra are regarded as phase I historical data. The simulation results are summarized in the following discussions.

Figure 6 shows a visual comparison between the denoised Raman spectra after wavelet shrinkage (Figure 6(a)) and those with generalized wavelet shrinkage (Figure 6(b)). Obviously, the peak intensities of denoised signals after generalized wavelet shrinkage (Figure 6(a)) are more uniform than peak intensities of denoised signals after wavelet shrinkage (Figure 6(b)).



(a) Wavelet Shrinkage (b) Generalized Wavelet Shrinkage

Figure 6. Comparison between wavelet shrinkage and generalized wavelet shrinkage

Quantitatively, we can compare the SNR of Raman spectra after wavelet shrinkage and generalized wavelet shrinkage. As for peak 1, P_1^* denotes the estimated peak intensity that describes signal amplitude; $\text{mean}(|P_1 - P_1^*|) + \text{std}(|P_1 - P_1^*|)$, represents that bias and standard deviation from the noise amplitude, and the calculation equation of SNR is

$$\text{SNR} = \frac{P_1^*/2}{\text{mean}(|P_1 - P_1^*|) + \text{std}(|P_1 - P_1^*|)}$$

According to Table 4, we can see that for peak 1 intensity, the S/N ratio is equal to 1.21 without wavelet denoising. Applying wavelet shrinkage method will increase the S/N ratio from 1.21 to 4.47. Moreover, the generalized wavelet shrinkage approach we proposed will increase the S/N ratio from 1.21 to 6.51.

Table 4. Comparison between wavelet shrinkage and generalized wavelet shrinkage based on peak 1

	λ	Mean $\lvert P_1 - P_1^* \rvert$	Std $\lvert P_1 - P_1^* \rvert$	SNR
Raw data	/	44.34	17.90	1.21
WS	160	8.49	8.48	4.45
	170	8.42	8.48	4.47
	180	8.08	8.38	4.58
GWS	0.48	6.70	4.90	6.51
	0.58	6.70	4.90	6.51
	0.68	6.70	4.90	6.51

* WS means "Wavelet Shrinkage"; GWS means "Generalized Wavelet Shrinkage"

Table 5. Comparison between wavelet shrinkage and generalized wavelet shrinkage based on peak 2

	λ	Mean $\lvert P_2 - P_2^* \rvert$	Std $\lvert P_2 - P_2^* \rvert$	SNR
Raw data	/	97.8082	52.7021	4.09
WS	260.00	47.79	35.35	7.40
	275.00	42.77	31.42	8.29
	290.00	40.04	28.28	9.00
GWS	0.14	24.77	19.11	14.02
	0.16	24.45	18.65	14.27
	0.18	25.12	18.87	13.98

According to Table 5, for peak 2, the S/N ratio can increase from 4.9 to 8.3 by wavelet shrinkage, and then to 14 by generalized wavelet shrinkage. Since the S/N ratio is correlated with the data acquisition time, the data acquisition time can be reduced without compromising the S/N ratio by this approach. This approach improves control accuracy, sensitivity and efficiency, and potentially increases the scalability of nanomanufacturing processes. Besides, the lambda can be selected under an area, which means the chosen penalized parameter is robust for the shrinkage.

2.5 Case Study: Investigation of In-Line Raman Spectroscopy for Buckypaper Quality Monitoring

In this case study, we explore the feasibility of using an in-line Raman spectroscopy on buckypaper samples to provide in situ data acquisition for quality control. The setup of

in-line Raman spectroscopy is shown in Figure 7. SWCNT and MWCNT buckypaper samples were fabricated and measured with the Raman spectroscopy. The acquisition time for static Raman spectroscopy varies from 0.01 seconds to 1 second for SWCNT buckypaper and from 0.1 seconds to 5 seconds for MWCNT buckypaper. In all measurements, Near Infra-Red (NIR) laser with a wavelength of 785nm and a power of 150mW was used to eliminate the effect of ambient light. A low magnification lens was used to achieve large focus tolerance.

Figure 7. Renishaw™ inVia micro-Raman system with custom-designed remote optical probe and roller sample stage

For every acquisition time, we measured 50 Raman spectra and processed data by the wavelet shrinkage and generalized wavelet shrinkage discussed in Section 2.3. The results for SWCNT buckypaper are given in Comparison between wavelet shrinkage and generalized wavelet shrinkage with Raman spectra on SWCNT buckypaper. According to Table 6, with the acquisition time as 1 second, the S/N ratio for peak 1 can be increased from 5.41 to 8.97 by wavelet shrinkage, 13.65 by generalized wavelet shrinkage respectively. The increases in the S/N ratio are 65.86 percent and 152.34 percent for peak 1. For peak 2, the S/N ratio increases from 17.86 to 37 by wavelet shrinkage and then 39.68 by generalized wavelet shrinkage. Similar conclusions can be drawn that with different

acquisition time, both wavelet shrinkage and generalized wavelet shrinkage can improve the S/N ratio, and generalized wavelet shrinkage improves more than conventional wavelet shrinkage.

Table 6. Comparison between wavelet shrinkage and generalized wavelet shrinkage with Raman spectra on SWCNT buckypaper

Exp time	/	Mean $\lvert P_1 - P_1^* \rvert$	Std $\lvert P_1 - P_1^* \rvert$	Mean $\lvert P_2 - P_2^* \rvert$	Std $\lvert P_2 - P_2^* \rvert$	Peak1 SNR	Peak2 SNR
T=1	Raw data	15.58	10.31	29.76	19.42	5.41	17.86
	WS	8.21	7.4	14.99	8.73	8.97	37.04
	GWS	5.99	4.27	14.01	8.13	13.65	39.68
T=0.5	Raw data	11.83	5.58	26.8	19.1	4.22	10.07
	WS	5.55	4.05	14.08	10.45	7.65	18.85
	GWS	5.04	3.77	12.06	7.27	8.33	23.92
T=0.1	Raw data	12.63	5.63	13.45	10.2	1.11	5.46
	WS	4.39	2.47	8.38	7.47	2.94	8.15
	GWS	3.9	2.93	8.18	6.03	2.96	9.09
T=0.05	Raw data	9.91	5.51	13.72	8.78	0.88	3.67
	WS	2.61	1.58	8.35	4.42	3.22	6.46
	GWS	2.53	1.58	6.44	4.01	3.28	7.89
T=0.01	Raw data	10.67	3.66	13.2	8.95	0.59	2.30
	WS	2.42	2.02	6.36	4.85	1.91	4.55
	GWS	2.18	1.61	5.74	4.38	2.24	5.04

* WS means "Wavelet Shrinkage"; GWS means "Generalized Wavelet Shrinkage";

In addition, the improvements of S/N ratios are different under different acquisition time. Based on the derivation of the generalized wavelet shrinkage, the performance of the generalized wavelet shrinkage will be similar to the conventional wavelet shrinkage if the noise has an insignificant dependence on the strength of the signal. Usually, a longer

acquisition time represents a higher intensity and stronger signal-noise dependence. In this situation, a generalized wavelet has a better performance than the conventional wavelet shrinkage.

Table 7. Comparison between wavelet shrinkage and generalized wavelet shrinkage with Raman spectra on MWCNT buckypaper

Exp time	/	Mean $\lvert P_1 - P_1^* \rvert$	Std $\lvert P_1 - P_1^* \rvert$	Mean $\lvert P_2 - P_2^* \rvert$	Std $\lvert P_2 - P_2^* \rvert$	Peak1 SNR	Peak2 SNR
T=5	Raw data	31.02	18.36	46.16	22.04	6.50	11.09
	WS	20.47	13.60	27.25	13.43	9.42	18.59
	GWS	16.17	12.04	24.58	6.63	11.37	24.23
T=1	Raw data	15.67	7.36	19.15	10.63	2.86	5.26
	WS	4.44	2.67	6.96	5.22	9.26	12.85
	GWS	4.44	2.67	6.75	4.14	9.26	14.38
T=0.5	Raw data	15.36	7.33	15.69	8.71	1.57	3.43
	WS	6.08	4.74	6.73	4.41	3.29	7.51
	GWS	3.93	2.38	6.85	3.64	5.65	7.97
T=0.2	Raw data	10.95	4.37	12.88	6.90	1.12	1.97
	WS	4.40	4.60	5.19	4.26	1.91	4.12
	GWS	2.45	2.20	4.35	2.87	3.69	5.41
T=0.1	Raw data	12.99	3.14	13.45	5.14	0.66	1.30
	WS	3.28	3.10	4.06	4.21	1.67	2.91
	GWS	1.96	2.05	3.38	2.66	2.67	3.99

The results for MWCNT buckypaper are shown in Table 7. With an acquisition time of 0.5 seconds for MWCNT buckypaper, the S/N ratio for peak 1 can rise from 1.57 to 3.29 by wavelet shrinkage, to 5.65 by generalized wavelet shrinkage respectively. The increased percentage on the S/N ratio is 109.55 percent and 259.87 percent for peak 1. For peak 2,

the S/N ratio increases from 3.43 to 7.51 by wavelet shrinkage and then 7.97 by generalized wavelet shrinkage.

According to the case study on SWCNT buckypaper and MWCNT buckypaper, the generalized wavelet shrinkage improves S/N ratio based on individual-dependent thresholding, which works better than the conventional wavelet shrinkage method. That means the generalized wavelet shrinkage approach will improve the accuracy and efficiency of quality control of the CNTs buckypaper manufacturing process.

2.6 Conclusion

Targeting the improvement of the S/N ratio and reducing data acquisition time for in-line Raman spectroscopy, this section focuses on developing a generalized wavelet shrinkage method to implement denoising of Raman spectrum that enables the use of Raman spectroscopy for in-line monitoring and control for the nanomanufacturing process. Several vital contents and conclusions are summarized as follows.

The noise in Raman spectra, including photon shot noise, sample-generated noise, instrument-generated noise, computationally generated noise, and externally generated noise are analyzed. The signal model of Raman spectrum is developed to describe the signal-noise dependence relationship. The model is validated by eight groups of experimental data with different types of CNTs buckypaper and different acquisition times. Results show that the value of linear parameters a and b remain stable and are not correlated with different materials and acquisition time. This allows for designing an adaptive denoising approach for individual wavelet coefficients.

Discrete wavelet decomposition can be implemented to transform Raman spectra signals into the wavelet domain. Based on the validated signal-noise dependence relationship, a novel generalized wavelet shrinkage approach is proposed to remove noise in each wavelet coefficient by individual adaptive wavelet thresholds, which outperforms the level-dependent conventional wavelet shrinkage method. The simulation and case study show the feasibility of the generalized wavelet shrinkage method. The technique can improve the S/N ratio dramatically or can be used to reduce data acquisition time without loss of S/N ratio, which lays a solid foundation for quick real-time monitoring and control of the nanomanufacturing process.

Moreover, the generalized wavelet shrinkage can be used to remove noise of signals with signal-dependent noise and realize better performance than the conventional wavelet shrinkage. Besides, it can also be applied to preprocess the calibration spectra with signal-dependent noise in order to prevent over-fitting and improve the accuracy in calibration modeling.

CHAPTER 3. A WAVELET-BASED PENALIZED MIXED-EFFECTS DECOMPOSITION FOR MULTICHANNEL PROFILE DETECTION OF IN-LINE RAMAN SPECTROSCOPY

Modeling and analysis of profiles, especially high-dimension nonlinear profiles, is an important and challenging topic in statistical process control. Conventional mixed-effects models have several limitations in multichannel profile detection of in-line Raman spectroscopy, such as inability to separate defect information from random effects, computational inefficiency, and inability to handle high-dimensional coefficients. In this chapter, a new wavelet-based penalized mixed-effects decomposition (PMD) method was proposed to solve the multichannel profile detection problem in Raman spectroscopy. The proposed PMD exploits a regularized high-dimensional regression with linear constraints to decompose the profiles into four parts: fixed effects, normal effects, defective effects, and signal-dependent noise. An optimization algorithm based on the accelerated proximal gradient (APG) was developed to do parameter estimation efficiently for the proposed model. Finally, the separated fixed effects coefficients, normal effects coefficients, and defective effects coefficients can be used to extract the quality features of fabrication consistency, within-sample uniformity, and defect information, respectively. Using a surrogated data analysis and a case study, we evaluated the performance of the proposed PMD method and demonstrated a better detection power with less computational time.

3.1 Introduction

Profile data, also called functional data or waveform data, has been commonly

observed in various industrial sensing and monitoring systems. Examples of such profile data include tonnage signals in the stamping process (Jin and Shi 2001), valve seat pressing data in the engine assembly process (Paynabar and Jin 2011), and so on. As a result, it is an important, yet challenging problem to model and analyze profile data for process monitoring and diagnosis. This problem is especially true in the continuous carbon nanotubes (CNTs) buckypaper manufacturing process, where in-line Raman spectroscopy instrument is used to provide Raman spectra, which is a special type of profile data, to indicate quality characteristics and properties of buckypaper.

CNTs buckypaper is an important multifunctional platform material with great potential for creating lightweight, high-performance materials for various applications due to its superior mechanical and electrical characteristics. One of the critical roadblocks to scale-up the production of high-quality buckypaper is on-line quality monitoring and control of its production system. As an effective off-line characterization method for nanostructure information, Raman spectroscopy has been widely used in batch-to-batch nanomanufacturing of CNTs buckypaper (Gommans et al. 2000; Raravikar et al. 2002). Recently, in-line Raman spectroscopy has been developed to monitor biochemical reactions and pharmaceutical crystallizations (Fevotte 2007), because of its non-destructive nature, fast data acquisition, and ability to provide detailed material information (Yue et al. 2017). However, the in-line Raman spectroscopy techniques of monitoring biochemical reactions and pharmaceutical crystallizations cannot be used to monitor nanomanufacturing of CNTs buckypaper majorly due to two reasons: (i) only specific intensity peaks are considered in the monitoring of biochemical reactions and pharmaceutical crystallizations, while in the monitoring of nanomanufacturing, the whole

profiles need to be analyzed; (ii) the techniques used in the monitoring of biochemical reactions and pharmaceutical crystallizations cannot separate defect information from the random effects. Thus, we focus on developing a profile modeling and detection approach based on in-line Raman spectroscopy for the roll-to-roll CNTs buckypaper fabrication process.

Raman analysis is critical for detecting the quality information of the product such as impurity, degree of chemical functionalization and alignment of CNTs in buckypaper. For example, Radial Breathing Mode (< 300 cm^{-1}) is used to determine the diameter of single-walled carbon nanotube (SWCNT); D-band (1250 – 1400 cm^{-1}) is related to the disorder or molecular defects in the CNTs structure; and D-band to G-band intensity ratio is an effective indicator of CNTs quality and functionalization (Cheng et al. 2010). In addition, polarized Raman provides angular dependence of the Raman intensity, and then the degree of CNTs alignment can be estimated (Jack et al. 2010).

A schematic diagram of in-line measurement procedure based on Raman spectroscopy is shown in Figure 8. In a roll-to-roll fabrication process of CNTs buckypaper, samples of Raman spectra are measured sequentially with Raman mapping technique. Each sample is shown as a red square in Figure 8. Within each sample, a specific number of Raman spectra are collected in the positions predetermined by design of experiment (DOE) to represent the quality information in the sample area. Usually, the maximin Latin Hypercube Design (Joseph and Hung 2008), which demonstrates good space-filling property and first-dimension projection property, is implemented to obtain the representative sampling positions, as shown in the drawing of partial enlargement in Figure 8. If each representative sampling position can be regarded as a sensor channel to collect

Raman spectra, the process modeling and detection for buckypaper fabrication process can be formulated as a multichannel profile modeling problem.

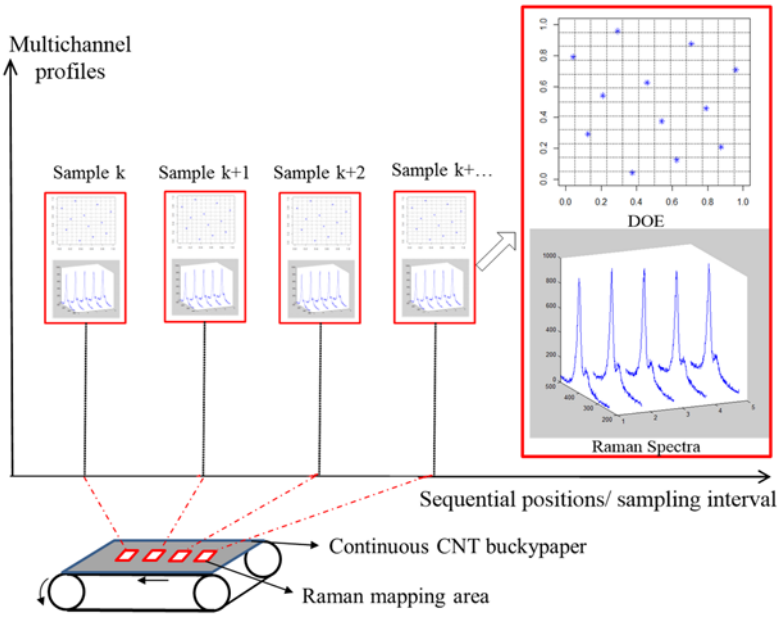


Figure 8. Schematic diagram of in-line measurement procedure based on Raman spectroscopy in a continuous CNTs buckypaper production process

Three quality characteristics are of critical importance during the statistical process control of the buckypaper fabrication process: (i) fabrication consistency of quality features measured by Raman spectroscopy, which indicates whether there is a gradual mean shift in the sequentially roll-to-roll fabrication of CNTs buckypaper; (ii) within-sample uniformity, which quantifies the uniformity of quality features (such as degree of alignment, degree of functionalization, nanotube distribution, and dispersion) in the sample area (shown as red square in Figure 8); (iii) within-sample defect information, which indicates whether there are defects in the CNTs buckypaper. Additionally, a specific band of Raman spectrum denotes corresponding defect information of the product. In our previous publication (Yue et al. 2017), the signal-dependent property of noise in Raman spectra has been identified and validated with the experimental data. A generalized wavelet

shrinkage method was proposed for denoising by applying individual adaptive wavelet thresholds. However, the denoised Raman spectra cannot be directly used in the multichannel profile monitoring because all components associated with different quality characteristics (e.g., fabrication consistency, uniformity, and defects) are all mixed in the signals. Therefore, profile modeling and signal decomposition for multi-channel signals are essential.

In order to monitor the aforementioned fabrication consistency, within-sample uniformity, and defect information for the in-line Raman spectroscopy, we propose a new multichannel profile modeling and analysis approach with consideration of the following two characteristics: (i) signal-dependent noise in Raman spectra, which has been validated in a previous study (Yue et al. 2017); (ii) multichannel profile features based on profiles in a sample reveals within-sample uniformity and defect information in the sample area, while changes among different samples illustrate the quality shift in the sequential long-term manufacturing process.

In this study, a new penalized mixed-effects decomposition (PMD) method is proposed for the multichannel profile detection in continuous nanomanufacturing process. The proposed PMD can separate the profiles into the following four parts: (i) fixed effects, which measure the fabrication consistency among samples over time; (ii) normal effects, which measure the within-sample uniformity of quality features; (iii) defective effects, which reveal the existence and location of the defects within a sample area; and (iv) signal-dependent noise, which is proportional to the Raman intensity in terms of the specific Raman shift.

To model waveform signals like Raman spectroscopy, wavelet basis is applied to generate design matrices in the proposed PMD. The merits of using wavelet basis will be discussed in Section 3.2. Another advantage is that random effects coefficients extracted by the PMD is sparse and has a much lower dimension than the original data due to the L_1 norm regularization term for defective effects coefficients. An accelerated proximal gradient (APG) algorithm is developed to effectively and efficiently estimate the parameters in the proposed PMD method. Sensitivity analysis is also conducted to determine the bounds of normal effects based on product tolerance.

The remainder of this chapter is organized as follows. Section 3.2 provides a literature review of linear and nonlinear profile modeling and analysis. Section 3.3 describes the proposed PMD formulation and compares it with the linear mixed-effects model and the smooth sparse decomposition method. In addition, an APG algorithm for parameter estimation is introduced, and a sensitivity analysis of the bounds is conducted. Section 3.4 presents a surrogated data analysis to test the performance of the PMD approach, followed with a real in-line Raman case study that is implemented in Section 3.5. Finally, a brief summary is discussed in Section 3.6.

3.2 Literature Review

This section provides a detailed literature review for profile modeling and analysis. Conventional methods of linear and nonlinear profile modeling and analysis are discussed first. Nonparametric techniques to handle non-smooth profiles are then introduced. Finally, mixed-effects model based profile modeling approaches and their limitations are discussed.

In the literature, most of the researchers focused on linear profile modeling and monitoring (Kang and Albin 2000; Kim et al. 2003; Mahmoud et al. 2007). Nonlinear profile monitoring was also investigated to address the nonlinearity of profiles in the different manufacturing processes. Different techniques, such as multiple and polynomial regression, cubic spline, and B-spline, were used to monitor nonlinear profile data (Mosesova et al. 2006; Zou et al. 2008; Jensen and Birch 2009; Shiau et al. 2009; Chang and Yadama 2010). However, the aforementioned approaches only focus on modeling of smooth and differentiable profiles, which are not applicable when the profiles include spikes or peaks such as those in the Raman spectroscopy of CNTs. To address this challenge, a piecewise method was proposed to divide a profile into several segments and fit them independently by using techniques mentioned above. However, these techniques require engineering background and estimation to decide all the breakpoints and segment intervals. Besides, different segmentation approaches also affect the modeling and monitoring results for a given application.

Some other techniques have been investigated to develop nonparametric methods to model non-smooth signals directly. Among all these techniques, principal components analysis (PCA), wavelet analysis, and empirical model decomposition (EMD) are the most common techniques that directly model nonlinear and nonstationary profiles.

PCA is a powerful unsupervised statistical procedure to reduce the dimension directly by orthogonal transformation such that the selected principal components account for most of the variability. Therefore, PCA is widely used in dimension reduction and statistical process control (Zhou and Jin 2005; Ding et al. 2006). However, there are some inherent limitations when the PCA method is used. One limitation is its weak

interpretability since all the random variables are included in the principal components. Another limitation is the inconsistency when applying PCA to process high-dimensional data with noise, proved by Johnstone and Lu (2012) by showing that the sample eigenvectors do not converge to the population eigenvectors and the noise does not average out if the dimension is much higher compared to the sample size. Sparse PCA was developed to improve the interpretability by imposing the sparsity constraint on the principal components (Zou et al. 2006). However, sparse principal components after dimension reduction usually do not cover all the bands in profiles, which leads to the loss of detection power in the band corresponding to the zero loadings. In addition, the defects tend to be correlated in the neighbor bands of profiles, and the sparse loadings ignore this kind of correlation.

Wavelet analysis is another powerful and widely applied nonparametric tool for time-frequency transformation because of the following three advantages: (i) it is an effective non-parametric approach that can model nonlinear profiles with sharp jumps and non-differentiable points; (ii) it provides multi-resolution, time-frequency and localized representation, which overcomes the limitations of traditional Fourier Transform (FT) and Short Time Fourier Transform (STFT); (iii) many natural images and signals have sparse representations using a wavelet basis, which is useful in the denoising and data compression. Due to the popularity of time-dependent nonlinear profile data in system monitoring and diagnosis, wavelet-based methods have been widely used to model the profile data for process modeling and monitoring problems (Jin and Shi 1999; Chicken et al. 2009; Paynabar and Jin 2011).

Additionally, EMD is a powerful time-frequency analysis technique. It can decompose the signal into a group of orthogonal components called intrinsic mode function (IMF) based on identification and interpolation of peaks and troughs in the signal. EMD has been widely utilized in various areas including process control, detection, and diagnosis (Huang et al. 1998; Lei et al. 2013). However, EMD suffers from three disadvantages: (i) the modes extracted by the method are not always physically meaningful, and the decomposition is limited to decompose each mode based on the varying frequency; (ii) it is noise sensitive, which is not suitable for Raman spectra with signal-dependent noise; (iii) it has mode mixing (identifiability) issue, which is defined as a single IMF including oscillations of dramatically disparate scales, or a component of a similar scale residing in different IMFs. Considering all these factors of PCA, wavelet analysis, and EMD, wavelet-based methods are preferred to model the profiles in the continuous nanomanufacturing.

In profile modeling and analysis, two kinds of variations are commonly considered. One is within-profile variation which is normally due to the random noise in the process such as environmental noise and measurement errors; the other is between-profile variation due to the inevitable profile-to-profile variability including process condition variations, fixture or tooling tolerance, and so on. Both within-profile variation and between-profile variation should be modeled to account for the total variabilities. This desire motivates to develop a mixed-effects model.

Researchers developed parametric or nonparametric mixed-effects model for profile monitoring. For example, parametric mixed-effects models were developed by treating a few model parameters as random effects (Mosesova et al. 2006; Jensen and Birch 2009), this kind of model is suitable for the data with complete domain knowledge and clear model

structure; A nonparametric mixed-effects model with B-spline and PCA was developed to monitor non-linear profiles (Shiau et al. 2009), which is usually achievable for smooth profiles. Paynabar and Jin used the wavelet-based mixed-effects model to characterize nonlinear profiles (Paynabar and Jin 2011). However, the existing mixed-effects model cannot be used directly to profile modeling and analysis in the continuous nanomanufacturing for the following three reasons: (i) A mixed-effects model can only measure between-profile variation by random effects, instead of separating normal between-profile variation and defective between-profile variation. The normal between-profile variation results from measurements and product uncertainty within the tolerance, while the defective between-profile variation is due to the defects in the buckypaper. (ii) There are too many coefficients used in a mixed-effects model, especially for high-dimensional profiles. This will result in both high false alarm rate and low detection rate in the multivariate process control. (iii) The algorithms to do parameter estimation of the mixed-effects model are not computationally efficient enough for in-line detection. Thus, we propose a PMD method that overcomes these three limitations from the conventional mixed-effects models. To address these challenges, the random effects are separated into normal effects and defective effects in our proposed model. Meanwhile, to reduce the dimensionality of the coefficients, wavelet basis and L_1 norm regularization is applied since it can achieve sparse representation of the data. The proposed methodology will be discussed in detail in the following section.

3.3 Penalized Mixed-effects Decomposition

In this section, we first review the conventional linear mixed-effects model (LMM) (Gałecki and Burzykowski 2013) and the smooth sparse decomposition (SSD) model (Yan

et al. 2017). We then introduce the penalized mixed-effects decomposition (PMD) for multichannel profile detection. After that, the PMD is compared with the LMM and the SSD. An accelerated proximal gradient (APG) algorithm is developed to estimate parameters of the PMD. Finally, a sensitivity analysis about the bounds of normal effects and the penalized parameter is discussed.

3.3.1 Review of the Mixed-effects Models and the Smooth Sparse Decomposition

A linear mixed-effects model (LMM) (Gałecki and Burzykowski 2013) is formulated for multichannel profile modeling as

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\varepsilon_i} \tag{18}$$

where $\boldsymbol{y_i}$ is a measurement profile; $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are design matrices associated with the fixed effects and the random effects, respectively. $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$ are coefficients vectors associated with the fixed effects and the random effects. $\boldsymbol{\varepsilon_i}$ is a noise vector that depends on the fixed effects and the random effects intensity because of the signal-dependent property.

The mixed-effects model has been widely used in the functional data analysis (Mosesova et al. 2006; Jensen and Birch 2009; Shiau et al. 2009; Paynabar and Jin 2011). However, this conventional mixed-effects model has the following limitations when it is applied to multichannel profile modeling based on in-line Raman spectra:

(i) It can only separate the fixed effects and the random effects, but cannot separate the normal random effects and the defective random effects. As a

consequence, it cannot be directly used in the monitoring of fabrication consistency, within-sample uniformity, and within-sample defect information for the Raman spectra simultaneously.

(ii) The coefficients extracted by a mixed-effects model are still high-dimensional, which brings challenges to achieve both a low false alarm rate and a high detection rate in the multivariate process monitoring problem.

(iii) In general, a restricted maximum likelihood (REML) method or an expectation-maximization (EM) method can be used to estimate the model parameters. However, those methods are computationally expensive, and they do not meet the need for real-time detection in the continuous roll-to-roll nanomanufacturing process, especially for profiles with heterogeneous noise, such as Raman spectra.

To separate the defect information from the smooth functional mean profile, a smooth sparse decomposition (SSD) method (Yan et al. 2017) was proposed to decompose each profile $\boldsymbol{y}$ into a smooth functional mean $\boldsymbol{B\theta}$, a sparse anomaly $\boldsymbol{B}_a\boldsymbol{\theta}_a$, and a noise vector $\boldsymbol{e}$ by solving the following optimization problem:

$$\begin{aligned} &\underset{\boldsymbol{\theta},\boldsymbol{\theta}_a}{\operatorname{argmin}} \left|\left|\boldsymbol{e}\right|\right|^2 + \lambda\boldsymbol{\theta}^T\boldsymbol{R}\boldsymbol{\theta} + \gamma|\boldsymbol{\theta}_a|_1 \\ &\text{Subject to. } \boldsymbol{y} = \boldsymbol{B\theta} + \boldsymbol{B}_a\boldsymbol{\theta}_a + \boldsymbol{e} \end{aligned} \tag{19}$$

where $||\cdot||$ and $|\cdot|_1$ are L_2 operator and L_1 norm operator, respectively. $\boldsymbol{B}$, $\boldsymbol{B}_a$, and $\boldsymbol{R}$, are design matrices for smooth part, sparse part, and a roughness matrix, respectively. λ and γ are tuning parameters to realize regularization for smooth coefficients $\boldsymbol{\theta}$ and sparse

coefficients $\boldsymbol{\theta}_a$. An accelerated proximal algorithm is applied to efficiently estimate $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_a$ simultaneously.

The SSD method is effective and efficient to separate the background, anomalies, and defect for both smooth functional data and images with the smooth background. However, it cannot be applied to the multichannel profile detection problem based on in-line Raman spectra for the following two reasons: (i) The SSD is designed for anomaly detection under smooth backgrounds, while Raman spectrum is inherently non-smooth. Therefore, a spline basis, which is used in the SSD, is not suitable for the non-smooth profiles. (ii) The loss function of the noise term does not consider the covariance matrix, which is not appropriate for signal-dependent noise in Raman spectra. Inspired by the SSD, we propose to decompose the defect effect and the normal effect by using wavelet basis to describe the waveform signal of Raman spectra and using a covariance matrix to consider the signal-dependent noise.

3.3.2 Penalized Mixed-effects Decomposition

A typical in-line Raman process monitoring problem is shown in Figure 8. The sequentially measured Raman profiles are grouped into N samples, and each sample has n_i profiles containing n_{ij} observations for different Raman shifts. The samples are indexed by $i = 1, \dots, N$ and the profiles in the ith sample are indexed by $j = 1, \dots, n_i$. The PMD for multichannel profile detection of Raman spectra is formulated as

$$\boldsymbol{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{W}\boldsymbol{\theta}_{ij} + \boldsymbol{W}_{\boldsymbol{a}}\boldsymbol{\delta}_{ij} + \boldsymbol{e}_{ij}$$

where $\boldsymbol{y}_{ij}$ is a measurement profile with dimension n, corresponding to the j^{th} profile in the i^{th} sample; $\boldsymbol{\mu}_i$ denotes fixed effects in the i^{th} sample; $\boldsymbol{W}$ and $\boldsymbol{W}_{\boldsymbol{a}}$ are wavelet-based design matrices with dimension $n \times p$ and $n \times q$ for normal effects and defective effects respectively; $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\delta}_{ij}$ are coefficients vectors associated with normal effects and defective effects with regards to the j^{th} profile in the i^{th} sample. $\boldsymbol{e}_{ij}$ represents a signal-dependent noise vector for the j^{th} profile in the i^{th} sample.

In the problem formulation, a weighted sum of square loss function is applied to model the signal-dependent noise and the L_1 norm regularization is implemented to encourage the sparsity of the defective coefficients. Aside from this, the normal effects will be bounded by $\boldsymbol{B}_L$ and $\boldsymbol{B}_U$, which can be determined by the product tolerance and a sensitivity analysis. Therefore, the proposed penalized mixed-effects decomposition (PMD) is formulated as:

$$\begin{gathered}\underset{\boldsymbol{\theta}_{ij},\boldsymbol{\delta}_{ij}}{\operatorname{argmin}}\ \boldsymbol{e}_{ij}^{T}\boldsymbol{\Omega}^{-1}\boldsymbol{e}_{ij}+\gamma\left|\boldsymbol{\delta}_{ij}\right|_1 \\ \text{subject to. } \boldsymbol{y}_{ij}=\boldsymbol{\mu}_i+\boldsymbol{W}\boldsymbol{\theta}_{ij}+\boldsymbol{W}_{\boldsymbol{a}}\boldsymbol{\delta}_{ij}+\boldsymbol{e}_{ij} \\ \boldsymbol{B}_L \preccurlyeq \boldsymbol{W}\boldsymbol{\theta}_{ij} \preccurlyeq \boldsymbol{B}_U \\ \boldsymbol{B}_U=\sqrt{\operatorname{diag}(\boldsymbol{\Omega})}\circ\boldsymbol{\omega}_1,\boldsymbol{B}_L=-\sqrt{\operatorname{diag}(\boldsymbol{\Omega})}\circ\boldsymbol{\omega}_2\end{gathered} \tag{20}$$

where $\boldsymbol{\Omega}$ is a variance-covariance matrix of Raman spectra noise; $\preccurlyeq$ denotes component-wise inequality; $\boldsymbol{B}_L$ and $\boldsymbol{B}_U$ are a lower bound and a upper bound for normal effects; $\operatorname{diag}(\boldsymbol{\Omega})$ denotes the vector of diagonal elements of $\boldsymbol{\Omega}$; $\circ$ represents the hadamard product (element-wise multiplication for two vectors). Other detailed introduction for each notation in Equation (20) can be illustrated as follows:

(i) $\boldsymbol{\mu}_i$ denotes the fixed effects within each sample. Monitoring $\boldsymbol{\mu}_i$ ($i = 1, \dots, N$) reveals the fabrication consistency of the samples, which is the long term mean shift among samples as the sequentially roll-to-roll CNTs buckypaper fabrication process goes on. Usually, it can be calculated by the median vector of multiple profiles in each sample for robust estimation.

(ii) $\boldsymbol{W\theta}_{ij}$ represents the normal effects within each sample. $\boldsymbol{\theta}_{ij}$ can be used to quantify the uniformity of quality features in the sample area, such as degree of alignment, degree of functionalization, nanotube distribution, dispersion, and so on. Because of the sparse representation property of most waveform signals in the wavelet analysis, the dimension of $\boldsymbol{\theta}_{ij}$ is much lower than the original dimension of Raman spectra, which can improve the pecision and efficiency of profile modeling. Additionally, $\boldsymbol{W\theta}_{ij}$ is bounded by $\boldsymbol{B}_U$ and $\boldsymbol{B}_L$. $\boldsymbol{B}_U$ and $\boldsymbol{B}_L$ determine the monitoring robustness for the uniformity. If the degree of uniformity requirement is strict, the bounds will be tight. We can estimate the bounds by the corresponding weighted standard deviations of the Raman intensity for a specific Raman shift. The weight vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ can be determined by the product tolerance and sensitivity analysis, which will be discussed in Section 3.3.5.

(iii) $\boldsymbol{W_a\delta}_{ij}$ describes the defective effects within each sample, which is to model the defect information. It reveals whether there are specific sampling points with defective quality features occurred in certain bands of a Raman spectrum. L_1 norm regularization encourages the sparsity of the anomalous coefficients, which synergizes with the inherent sparsity of wavelet coefficients. Penalized parameter γ

can be determined by the cross validation described in Section 3.3.6. Furthermore, according to the mathematical link between wavelet coefficients and original signal segments (Jin and Shi 1999), we can derive which specific bands of a Raman spectrum have defects based on the location of the non-zero coefficients.

(iv) $\boldsymbol{e}_{ij}$ denotes the signal-dependent noise. The detailed noise source and its signal-dependent characteristic are validated in the paper (Yue et al. 2017). A weighted sum of square loss function for the signal-dependent noise can separate the noise accurately, which will improve the effect and efficiency of profile monitoring based on the random effect. The variance-covariance matrix can be obtained by the real Raman spectra data collected during phase I analysis (Yue et al. 2017).

(v) $\boldsymbol{W}$ and $\boldsymbol{W_a}$ are the wavelet-based design matrices for the normal effects and the defective effects, which are determined by the selected wavelet basis and its coefficients. $\boldsymbol{W}$ is required to be orthogonal. The wavelet basis usually can be selected upon a defined criterion (Yue et al. 2017). It is worth noting that similar to the design matrices determination in the mixed-effects model (Gałecki and Burzykowski 2013), usually, $\boldsymbol{W_a}$ is chosen as a subset of columns of $\boldsymbol{W}$ based on random effects pattern from engineering knowledge. If we do not know the design matrix columns associated with random effects, we can choose $\boldsymbol{W_a}$ equals to $\boldsymbol{W}$. Even $\boldsymbol{W}$ and $\boldsymbol{W_a}$ are the same, the identifiability can also be ensured. We realize the identifiability of coefficients by two settings: (1) the $\boldsymbol{W\theta}_{ij}$ is bounded by $\boldsymbol{B}_U$ and $\boldsymbol{B}_L$; (2) both $\boldsymbol{e}_{ij}^T\boldsymbol{\Omega}^{-1}\boldsymbol{e}_{ij}$ and $\gamma\left|\boldsymbol{\delta}_{ij}\right|_1$ are considered in the loss function. For example, firstly, $\boldsymbol{\mu}_i$ can be isolated by the median profile. Since $\boldsymbol{\delta}_{ij}$ is in the loss function,

while $\boldsymbol{\theta}_{ij}$ is not in the loss function. The model will distribute as many components as possible into $\boldsymbol{\theta}_{ij}$ until $\boldsymbol{W\theta}_{ij}$ is bounded. If $\boldsymbol{W\theta}_{ij}$ is bounded, other components in the profiles will be distributed into $\boldsymbol{\delta}_{ij}$ and $\boldsymbol{e}_{ij}$ according to the trade-off between $\boldsymbol{e}_{ij}^{T}\boldsymbol{\Omega}^{-1}\boldsymbol{e}_{ij}$ and $\left|\boldsymbol{\delta}_{ij}\right|_{1}$.

3.3.3 Comparison between the PMD and the LMM/SSD

As shown in Figure 9, the PMD is compared with the conventional LMM and the SSD. In comparison to the LMM, fixed effects obtained by the PMD correspond to fixed effects separated by the LMM, while random effects in the LMM can be further decomposed into normal effects and defective effects by the PMD. The normal effects are generated from the uncertainty of the spectra measurement and normal production conditions. Usually, the tolerance of product design is used to control the uncertainty of manufacturing and measurement process. However, defective effects are generated from defects such as impurities mixed in the CNTs buckypaper or large interspace hole in the aligned CNTs. Thus, the PMD could be regarded as an improved mixed-effects model that can achieve more detailed decomposition.

The SSD decomposes each profile $\boldsymbol{y}$ into a smooth functional mean $\boldsymbol{B\theta}$, a sparse anomaly $\boldsymbol{B}_{a}\boldsymbol{\theta}_{a}$, and a noise vector $\boldsymbol{e}$, which works well for the data with smooth background and sparse defects. However, when applied to the non-smooth Raman spectra, the defective effects can also be very smooth like the other band (e.g. band between 200-300 of Raman shift index in Figure 10), not associated with D-band or G-band, in a Raman spectrum. In this situation, the defect information will be mistakenly regarded as the smooth background in the SSD. On the other hand, the sparse peak regarded as an anomaly

in the SSD may also be the fixed effects or normal effects, such as the G-band peak in a Raman spectrum. The PMD combines the advantage of the LMM and the SSD and decomposes a profile into fixed effects, normal effects, defective effects, and signal-dependent noise, which can be used to monitoring different quality features like shift in fixed effects, degree of uniformity, and defects.

In summary, Figure 9 provides a schematic diagram of decomposition comparison between the PMD and the LMM/SSD. Intuitively, two information is conveyed by the figure: (i) the PMD can realize deeper decomposition than the LMM; (ii) the PMD and the SSD decompose the profiles from different point-of-views.

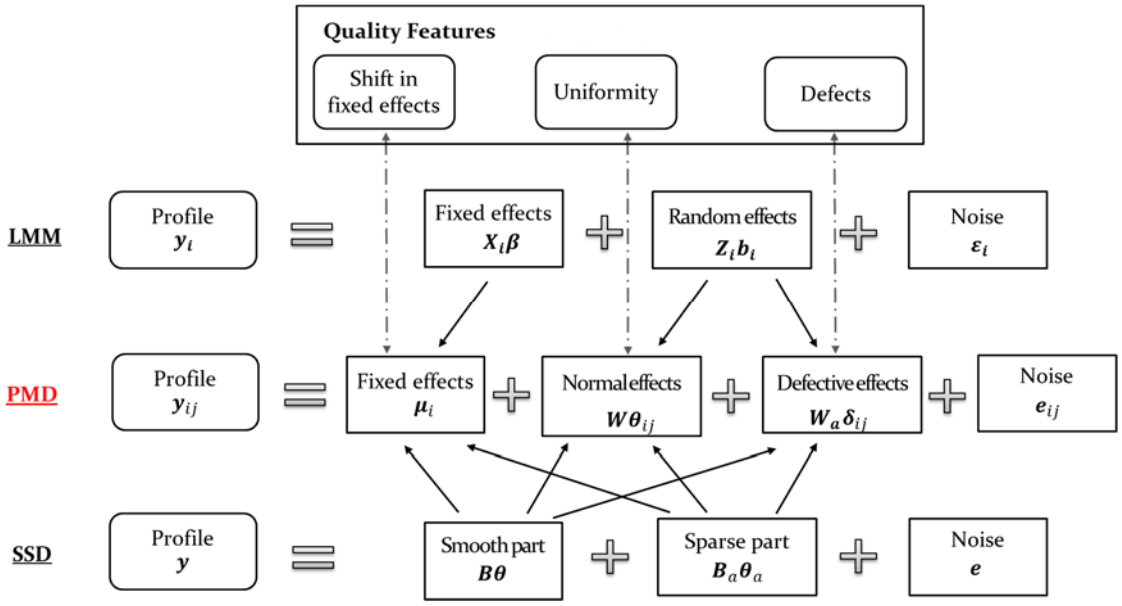


Figure 9. Schematic diagram of decomposition comparison among the PMD, the LMM, and the SSD

3.3.4 Algorithm for Parameter Estimation

In this section, the PMD model is simplified to a weighted LASSO (Zou 2006) problem. After that, the APG algorithm to solve this convex optimization problem is derived.

It can be shown that the variance-covariance matrix $\boldsymbol{\Omega}$ is a diagonal matrix with all the diagonal entries equal to the variance of Raman intensity corresponding to the related Raman shift. Thus, $\boldsymbol{\Omega}$ is invertible and each diagonal entry is positive. Equation (20) can be transformed into Equation (21).

$$\begin{aligned} \min_{\boldsymbol{\theta}_{ij},\boldsymbol{\delta}_{ij}} \ & (\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}\boldsymbol{\theta}_{ij} - \boldsymbol{W}_{\boldsymbol{a}}\boldsymbol{\delta}_{ij})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}\boldsymbol{\theta}_{ij} - \boldsymbol{W}_{\boldsymbol{a}}\boldsymbol{\delta}_{ij}) \\ & + \gamma \left|\boldsymbol{\delta}_{ij}\right|_1 \\ & \text{subject to.} \quad \boldsymbol{B}_L \preccurlyeq \boldsymbol{W}\boldsymbol{\theta}_{ij} \preccurlyeq \boldsymbol{B}_U \end{aligned} \tag{21}$$

$$\boldsymbol{B}_U = \sqrt{diag(\boldsymbol{\Omega})} \circ \boldsymbol{\omega}_1, \qquad \boldsymbol{B}_L = -\sqrt{diag(\boldsymbol{\Omega})} \circ \boldsymbol{\omega}_2$$

Thus, Equation (21) is a weighted LASSO problem with linear constraints. We can further prove the penalized mixed-effects decomposition shown in Equation (21) is equivalent to a weighted LASSO problem, and a detailed description is shown in Proposition 1.

Proposition 1. Assume $\boldsymbol{W}$ is orthogonal, the PMD in Equation (21) is equivalent to minimize a weighted sum of square loss function plus L_1 norm regularization in the form of

$$\operatorname*{argmin}_{\boldsymbol{\delta}_{ij}} F(\boldsymbol{\delta}_{ij}) := (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}^*)^T \boldsymbol{\Omega}^{-1} (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}^*) + \gamma \left|\boldsymbol{\delta}_{ij}\right|_1 \tag{22}$$

where $\boldsymbol{\theta}_{ij}^{*} = \boldsymbol{W}^{T}\boldsymbol{z}_{ij} \circ \boldsymbol{I}\left(\boldsymbol{B}_{L} \preccurlyeq \boldsymbol{z}_{ij} \preccurlyeq \boldsymbol{B}_{U}\right) + \boldsymbol{W}^{T}\boldsymbol{B}_{L} \circ \boldsymbol{I}\left(\boldsymbol{z}_{ij} \preccurlyeq \boldsymbol{B}_{L}\right) + \boldsymbol{W}^{T}\boldsymbol{B}_{U} \circ \boldsymbol{I}\left(\boldsymbol{z}_{ij} \succcurlyeq \boldsymbol{B}_{L}\right)$, $\boldsymbol{z}_{ij} := \boldsymbol{y}_{ij} - \boldsymbol{\mu}_{i} - \boldsymbol{W}_{\boldsymbol{a}}\boldsymbol{\delta}_{ij}$, $\boldsymbol{I}(\cdot)$ is an indicator operator. The formulation above is the weighted LASSO formulation.

The proof can be found in Appendix A.1.

Conventional LASSO solvers, such as least angle regression (LARS), cannot efficiently solve the weighted LASSO problem for high-dimensional data (Yan et al. 2017). Consequently, an effective and efficient algorithm should be developed to solve the problem to satisfy the in-line monitoring requirement. A popular efficient optimization algorithm, the proximal gradient method (Parikh and Boyd 2014) can be used to solve the Equation (22). The proximal gradient method is focused on minimization of summation of a group of convex function. The function in Equation (22) can be regarded as two sub-functions: $f\left(\boldsymbol{\delta}_{ij}\right) := \left(\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*}\right)^{T}\boldsymbol{\Omega}^{-1}\left(\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*}\right)$ and $g\left(\boldsymbol{\delta}_{ij}\right) := \gamma\left|\boldsymbol{\delta}_{ij}\right|_{1}$. Since $\boldsymbol{\Omega}^{-1}$ is a positive semi-definite matrix, $f\left(\boldsymbol{\delta}_{ij}\right)$ is a convex differentiable function. Besides, $g\left(\boldsymbol{\delta}_{ij}\right)$ is a non-differentiable convex function. We also prove in Proposition 2 that $f(\cdot)$ is Lipchiz continuous, which can guarantee the proximal gradient method to converge.

Proposition 2. $f(\cdot)$ is Lipchiz continuous, which implies there is a constant L that makes gradient function $\nabla f(\cdot)$ satisfy $\|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\| \leq L\|\boldsymbol{a} - \boldsymbol{b}\|$ for any $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{R}$ with $L = \frac{2}{\min(\mathrm{diag}(\boldsymbol{\Omega}))}\|\boldsymbol{W}_{\boldsymbol{a}}\|_{2}^{2}$, where $\|\cdot\|_{2}^{2}$ denotes square of matrix spectral norm.

The proof can be found in Appendix A.2.

Therefore, the proximal gradient algorithm can optimize $F(\boldsymbol{\delta}_{ij})$ by using an iterative algorithm given by

$$\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k)} = \underset{\boldsymbol{\delta}_{ij}}{\operatorname{argmin}} \left\{ f\left(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}\right) + \langle \boldsymbol{\delta}_{ij} - \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}, \nabla f\left(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}\right) \rangle + \frac{L}{2} \left\| \boldsymbol{\delta}_{ij} - \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} \right\|^2 + \gamma \left| \boldsymbol{\delta}_{ij} \right|_1 \right\}$$

where super-indices (k) and $(k-1)$ denote iteration numbers and $\langle \cdot, \cdot \rangle$ means the inner product operator.

Proposition 3. The proximal gradient algorithm for the PMD in Equation (22), given by $\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k)} = \underset{\boldsymbol{\delta}_{ij}}{\operatorname{argmin}} \left\{ f\left(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}\right) + \langle \boldsymbol{\delta}_{ij} - \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}, \nabla f\left(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}\right) \rangle + \frac{L}{2} \left\| \boldsymbol{\delta}_{ij} - \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} \right\|^2 + \gamma \left| \boldsymbol{\delta}_{ij} \right|_1 \right\}$, has a closed-form solution for each iteration k, in the form of a soft-thresholding function as follows:

$$\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k)} = S_{\frac{\gamma}{L}}\left(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} + \frac{2}{L} \boldsymbol{W}_{\boldsymbol{a}}^{\boldsymbol{T}} \boldsymbol{\Omega}^{-1} \left(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*} - \boldsymbol{W}_{\boldsymbol{a}} \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} \right)\right)$$

with $L = \frac{2}{\min(\operatorname{diag}(\boldsymbol{\Omega}))} \|\boldsymbol{W}_{\boldsymbol{a}}\|_2^2$.

$S_{\frac{\gamma}{L}}(x) = \operatorname{sgn}(x)(|x| - \frac{\gamma}{L})_+$ is the soft-thresholding operator, and $\operatorname{sgn}(\mathrm{x})$ is the sign function and $\mathrm{x}_+ = \max(x, 0)$. The proof of Proposition 3 can be found in Appendix A.3.

Therefore, the convergence of the proximal gradient algorithm is guaranteed (Parikh and Boyd 2014). In order to improve the convergence speed of the optimization algorithm, an accelerated proximal gradient method (Tseng 2008) can be applied to adjust

the original proximal gradient algorithm. The final algorithm can be summarized as follows.

Algorithm 1: Accelerated proximal gradient based algorithm for the penalized mixed-effects decomposition

While $\boldsymbol{j = 1 : N}$

Initialization:

$\boldsymbol{\delta_{ij}^{(0)} = 0}$, $\boldsymbol{L = \frac{2}{\min(\mathrm{diag}(\Omega))}\,\|W_a\|_2^2}$, $\boldsymbol{\mu_i = \mathrm{median}(y_{i\cdot})}$, $\boldsymbol{x^{(0)} = 0}$, $\boldsymbol{t_0 = 0}$

End

While $\boldsymbol{\left|\delta_{ij}^{(k-1)} - \delta_{ij}^{(k)}\right| > \epsilon}$

Let $\boldsymbol{\theta_{ij}^{*} = W^T z_{ij} \circ I\left(B_L \preccurlyeq z_{ij} \preccurlyeq B_U\right) + W^T B_L \circ I\left(z_{ij} \preccurlyeq B_L\right) + W^T B_U \circ I\left(z_{ij} \succcurlyeq B_L\right)}$

Update $\boldsymbol{\delta_{ij}^{(k)} = S_{\frac{\gamma}{L}}\left(x^{(k-1)} + \frac{2}{L} W_a^T \Omega^{-1}\left(y_{ij} - \mu_i - W\theta_{ij}^{*} - W_a x^{(k-1)}\right)\right)}$

Update $\boldsymbol{t_k = \left(1 + \sqrt{1 + 4t_{k-1}^2}\right)/2}$

Update $\boldsymbol{x^{(k)} = \delta_{ij}^{(k-1)} + \frac{t_{k-1}-1}{t_k}\left(\delta_{ij}^{(k-1)} - \delta_{ij}^{(k-2)}\right)}$

End

End

3.3.5 Sensitivity Analysis of the Bounds

In the modeling and analysis, how to determine the bounds $\boldsymbol{B}_L$ and $\boldsymbol{B}_U$ is a very important topic. The bounds are related to the product uniformity tolerance, which is a tradeoff between manufacturing cost and product quality. A sensitivity analysis is conducted to study how the optimal value changes with respect to perturbations of the constraints.

According to the analysis in Section 3.3.4, the optimization Equation (21) is convex. Meanwhile, Slater's condition holds for linear box constraints in this optimization problem (Boyd and Vandenberghe 2004). Thus, the strong duality does hold for this

optimization problem, which indicates that the best bound that can be obtained from the Lagrange dual function is tight.

In order to investigate the sensitivity of the optimal value with respect to the perturbations of the constraints, a perturbed version of the optimization is formulated as

$$\begin{aligned} \min_{\boldsymbol{\theta}_{ij},\boldsymbol{\delta}_{ij}} \ & (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}) + \gamma \left|\boldsymbol{\delta}_{ij}\right|_1 \\ \text{Subject to.} \quad & \boldsymbol{B}_L - \boldsymbol{W}\boldsymbol{\theta}_{ij} \preccurlyeq \Delta\boldsymbol{B}_L \\ & \boldsymbol{W}\boldsymbol{\theta}_{ij} - \boldsymbol{B}_U \preccurlyeq \Delta\boldsymbol{B}_U \end{aligned} \tag{23}$$

where $\Delta\boldsymbol{B}_L$ and $\Delta\boldsymbol{B}_U$ are perturbated values of the bounds of normal effects, and $\boldsymbol{z}_{ij} = \boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}_a\boldsymbol{\delta}_{ij}$.

Let $G(\boldsymbol{\theta}_{ij}, \boldsymbol{\delta}_{ij}) := (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}) + \gamma \left|\boldsymbol{\delta}_{ij}\right|_1$. Let $p^*(\Delta\boldsymbol{B}_L, \Delta\boldsymbol{B}_U)$ be the optimal value of the perturbed problem in Equation (23). Thus, $p^*(0,0)$ is the optimal value of the original unperturbed problem in Equation (21). Let $(\boldsymbol{\lambda_1^*}, \boldsymbol{\lambda_2^*})$ be optimal for the dual equation, then we obtain (Boyd and Vandenberghe 2014):

$$p^*(\Delta\boldsymbol{B}_L, \Delta\boldsymbol{B}_U) \geq p^*(0,0) - \boldsymbol{\lambda_1^{*T}}\Delta\boldsymbol{B}_L - \boldsymbol{\lambda_2^{*T}}\Delta\boldsymbol{B}_U$$

We can draw the conclusions that (i) if λ_{1k}^* or λ_{2k}^* is large and we tighten the k^{th} constraint by choosing the negative $\Delta\boldsymbol{B}_{Lk}$ or $\Delta\boldsymbol{B}_{Uk}$, then the optimal value $p^*(\Delta\boldsymbol{B}_L, \Delta\boldsymbol{B}_U)$ will increase greatly. (ii) if λ_{1k}^* or λ_{2k}^* is small and we lossen the k^{th} constraint by choosing the positive $\Delta\boldsymbol{B}_{Lk}$ or $\Delta\boldsymbol{B}_{Uk}$, then the optimal value $p^*(\Delta\boldsymbol{B}_L, \Delta\boldsymbol{B}_U)$ will not decrease too much. Based on the Karush–Kuhn–Tucker (KKT) condition of the optimality (Boyd and Vandenberghe 2014), the constraint $\boldsymbol{B}_{Lk} = (\boldsymbol{W}\boldsymbol{\theta}_{ij})_k$ will be active if λ_{1k}^* is non-zero. In

practice, the selection of bounds can be done by running the modeling and analysis based on phase-I datasets, and checking whether the detected defective part is acceptable under the corresponding uniformity tolerance. If some small peaks detected as defective parts are acceptable under uniformity tolerance, we can loose the corresponding bounds to realize better decomposition. The sensitivity analysis of the bounds will be discussed in the surrogated data analysis in the next section.

3.3.6 Criterion for Selection of the Penalized Parameter

The penalized parameter γ directly controls the trade-off of weighted sum of square loss function for the signal-dependent noise $\boldsymbol{e}_{ij}^T\boldsymbol{\Omega}^{-1}\boldsymbol{e}_{ij}$ and L_1 norm regularization of the anomalous coefficients $\left|\boldsymbol{\delta}_{ij}\right|_1$. Thus, γ has a significant impact on the performance of the PMD. In our study, we use the cross validation to determine the value of the penalized parameter γ. The optimal penalized parameter γ suffices the minimum mean sum of square of the difference between the original signals and estimated signals.

$$\gamma^* = \underset{\gamma \in R^+}{\operatorname{argmin}} \frac{1}{n(\sum_{i=1}^{N} n_i)} \sum_{i=1}^{N} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \bar{\bar{\boldsymbol{y}}}_{i[j]})^T (\boldsymbol{y}_{ij} - \bar{\bar{\boldsymbol{y}}}_{i[j]})$$

where $\boldsymbol{y}_{ij}$ is the original j^{th} measurement profile with dimension n in the i^{th} sample; $\bar{\bar{\boldsymbol{y}}}_{\boldsymbol{i}[\boldsymbol{j}]}$ is the estimated j^{th} measurement profile in the i^{th} sample. $\bar{\bar{\boldsymbol{y}}}_{\boldsymbol{i}[\boldsymbol{j}]}$ is estimated by the mean profile of other reconstructed profiles except the j^{th} one in the i^{th} sample. $\bar{\bar{\boldsymbol{y}}}_{\boldsymbol{i}[\boldsymbol{j}]} = \sum_{h=1,h\neq j}^{n_i} (\widehat{\boldsymbol{\mu}}_i + \boldsymbol{W}\widehat{\boldsymbol{\theta}}_{ih} + \boldsymbol{W}_{\boldsymbol{a}}\widehat{\boldsymbol{\delta}}_{ih})$.

This criterion makes use of the uniformity of different profiles in the same sample. In-control profiles should be used when the optimal penalized parameter γ^* is selected.

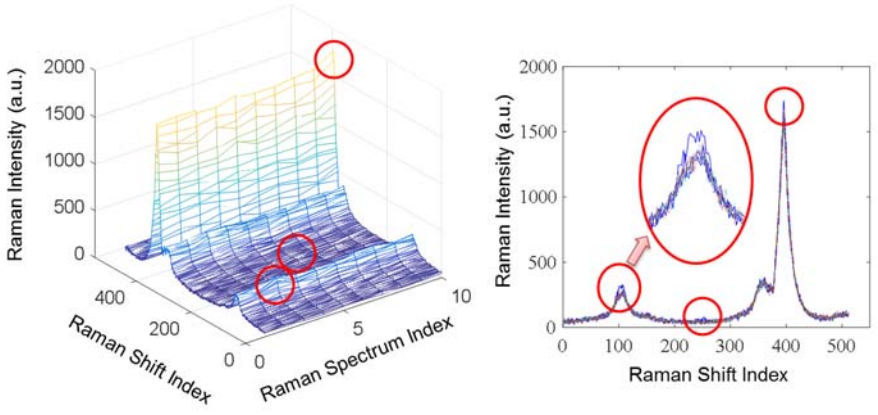
3.4 Surrogated Data Analysis of Raman Spectra

In this section, the performance of the PMD is evaluated through the surrogated Raman spectra from a practical experiment. Twenty Raman spectra are collected by measuring a piece of single-wall CNTs buckypaper sample. A Renishaw InVia Confocal micro-Raman Microscopy is used in the experiment, which is set with 785-nm laser source and 0.5-second exposure time. Ensemble averaging is applied to improve the signal-to-noise ratio and generate the mean vector as the ideal Raman spectrum for further surrogate profiles generation. Ten surrogated Raman spectra are generated based on the ideal Raman spectrum. Specifically, we consider three types of defects in different bands of the profiles: D-band in the third profile, a middle band between D-band and G-band in the 6th profile, and G-band in the 10th profile, respectively. Moreover, signal-dependent noise is generated by $\boldsymbol{e}\sim\boldsymbol{N}(\mathbf{0}, \text{diagonal}(\boldsymbol{S}) + 42)$, and $\boldsymbol{S}$ is the corresponding spectrum with dimension 512. The $\text{diagonal}(\boldsymbol{S})$ denotes the matrix with diagonal vector $\boldsymbol{S}$ and zero non-diagonal elements. The signal-dependent noise distribution has been validated by the real Raman spectra data in the paper (Yue et al. 2017). All the profiles with defect information and signal-dependent noise are shown in Figure 10 and the defective parts are marked by the red circles. We can observe that the defects are hidden in the noise, and it is not feasisble to pick up defective profiles mannually.

3.4.1 Performance Comparison between the PMD and the LMM/SSD

We compare our proposed PMD with two benchmark methods: the LMM and the SSD as introduced in Section 3.3. For the LMM, the median vector of all the profiles are extended to generate the design matrix of the fixed effects; and the design matrix of the

random effects are generated based on the linear spline regression settings (Gałecki and Burzykowski 2013). For the SSD, two cubic spline basis matrices are generated as the design matrices with 100 and 128 knots; and the tuning parameters are selected automatically based on the generalized cross validation criterion and Otsu's method (Yan et al. 2017). The SSD is implemented to analyze all ten profiles. In the PMD method, the penalized parameter γ can be determined by the cross validation in Section 3.3.6, and the selected penalized parameter γ^* is 0.15. The weight vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ were adjusted carefully to make that the sensitivity of detecting defects in the D-band and the G-band regions is higher than other regions since these regions are closely related to the degree of functionalization and degree of alignment of CNTs buckypaper. According to the comparison among Coiflet, Daubechies, and Symmlet families, Daubechies 18 is chosen as the wavelet basis upon a criterion defined by the optimal estimated peak intensity (Yue et al. 2017). Wavelet design matrices $\boldsymbol{W}$ and $\boldsymbol{W_a}$ are generated based on the first level of wavelet coefficients. The variance-covariance matrix is $\boldsymbol{\Omega} = \text{diagonal}(\boldsymbol{S}) + 42$.



(a) 3-D mesh plot of ten Raman spectra (b) 2-D plot of ten Raman spectra

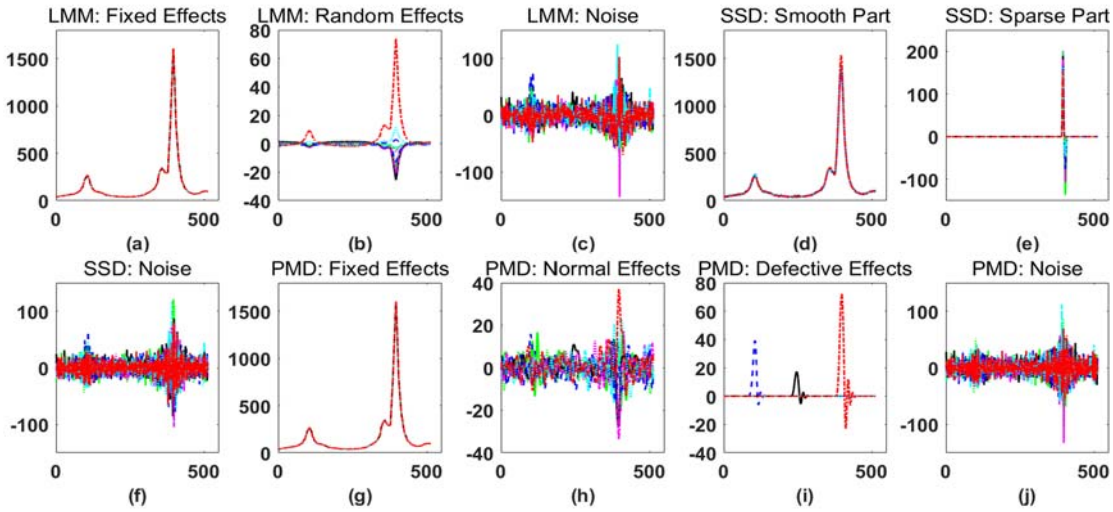
(Note: defects are marked by the red circles)

Figure 10. Ten Raman spectra with three defects in different band of different profiles

In order to evaluate the performance of the proposed PMD and the benchmark methods, we have defined the following criteria for comparison: (i) detection rate (DR), defined as the proportion of the real defect points detected as the defect points,; (ii) false alarm rate (FAR), defined as the percentage of the false classification of defect points in all the non-defective points; (iii) detected peak intensity difference (DPID), defined as $1/n_d \sum_{i=1}^{n_d} |\max(I_i) - \max(I_i^*)|/|\max(I_i^*)|$, where n_d denotes the number of real defect bands, $\max(I_i)$ denotes the detected peak intensity in band i, $\max(I_i^*)$ denotes the real peak intensity in the band I; (iv) the mean square error (MSE); and (v) the computation time. (i) is the larger the better. (ii-v) is the smaller the better.

The decomposition results by the LMM, the SSD, and the PMD are shown in Figure 11 and Table 8. The fixed effects, random effects, and noise separated by the LMM are shown in Figure 11(a-c), respectively. We can observe that the random effects separated by the LMM include both normal and defective effects. Also, part of defective effects remains in the noise. The detection rate is 100.00%, but the false alarm rate is as high as 85.69%, and the detected peak intensity difference is 77.73%. Thus, it is difficult to quantify the uniformity and detect the defects based on the LMM. Figure 11(d-f) shows the decomposition results by the SSD, including smooth part, sparse part, and noise. The false alarm rate is only 2.21% but the detection rate is very low (13.33%), and the detected peak intensity difference is huge (74.10%). The reason for such a low detection rate and a large detected peak intensity difference is that the SSD only applies to curves with smooth functional mean, and it mistakenly detects the spikes existed in the G-band of Raman spectra in the normal situation as anomalies due to the spline basis, while the simulated

defects remain in the smooth part with fixed effects and normal effects. Therefore, we cannot realize the quantification of uniformity and detection only based on the smooth part.



(Note: all the x axis denotes Raman shift index, all the y axis denotes the Raman intensity (a.u.))

Figure 11. Decomposition comparison among the LMM, the SSD and the PMD

With the PMD, the result of successfully separating the whole profiles into four parts including fixed effects within each sample, normal effects within each sample, defective effects within each sample, and signal-dependent noise are shown in Figure 11(g-j). Monitoring the fixed effects in Figure 11(g) for different samples can reveal fabrication consistency of quality features, which is the long-term mean shift among samples as the sequentially buckypaper fabrication process goes on. Meanwhile, the normal effects, which is usually due to the inherent uncertainty of regular product fabrication and measurement, is separated and shown in Figure 11(h). As shown in Figure 11(i), the defects information in the 3rd, 6th and 10th profiles, can be detected in the defective effects without false detection in other profiles. The detection rate is 100.00%. The false alarm rate is 2.63% because of the boundary effect when applying wavelet basis as shown in Figure 11(i). The detected peak intensity difference is 30.95%. In addition, the signal-dependent noise is

shown in Figure 11(j). According to these four decomposed effects, we can monitor fabrication consistency, within-sample uniformity, and within-sample defect information of quality features based on in-line Raman spectroscopy.

Table 8. Detect rate, false alarm rate, detected intensity difference and computation time for different methods

	LMM	SSD	PMD
Detection Rate (DR)	100.00%	13.33%	100.00%
False Alarm Rate (FAR)	85.69%	2.21%	2.63%
Detected Peak Intensity Difference (DPID)	77.73%	74.10%	30.95%
Mean Square Error (MSE)	9.93	10.06	9.18
Computation Time	1203.27s	0.55s	0.94s

Table 8 shows that the mean square error (MSE) of the LMM, the SSD, and the PMD are 9.93, 10.06, and 9.18, respectively. The MSE of the PMD is the smallest one, which makes sense because the normal effects and defective effects are separated thoroughly. The computation time of the LMM, SDD and PMD are 1203.27 seconds, 0.55 seconds and 0.94 seconds. The computation time of the LMM is very large since the estimation based on restricted maximum likelihood estimation is very slow. While the accelerated proximal gradient algorithm for the SSD and the PMD is very quick. The computation efficiency of the SSD and the PMD meet the requirements of real-time process monitoring based on in-line Raman spectroscopy.

The dimension of the original Raman spectra is 512 while only 64 wavelet coefficients in the first level are extracted for normal effects. The dimension of $\boldsymbol{\theta}_{ij}$ is much lower than the original dimension of Raman spectra, which improves the efficacy and efficiency of profile modeling. Similarity measure can then be used to quantify the

uniformity among ten profiles. Another 64 wavelet coefficients of defective effects $\boldsymbol{\delta}_{ij}$ includes the defect information in the corresponding profiles. Due to the property of wavelet transformation, each defective coefficient is related to a specific band in the Raman spectra, which provides useful information for root cause analysis of the defect.

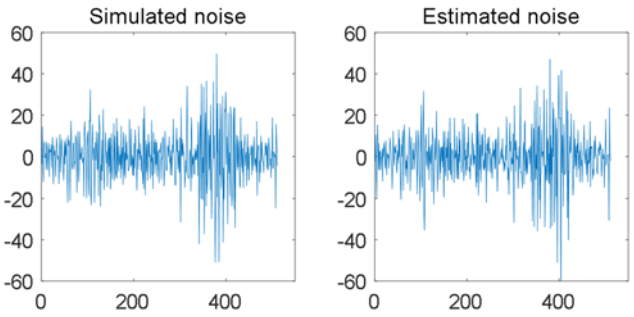


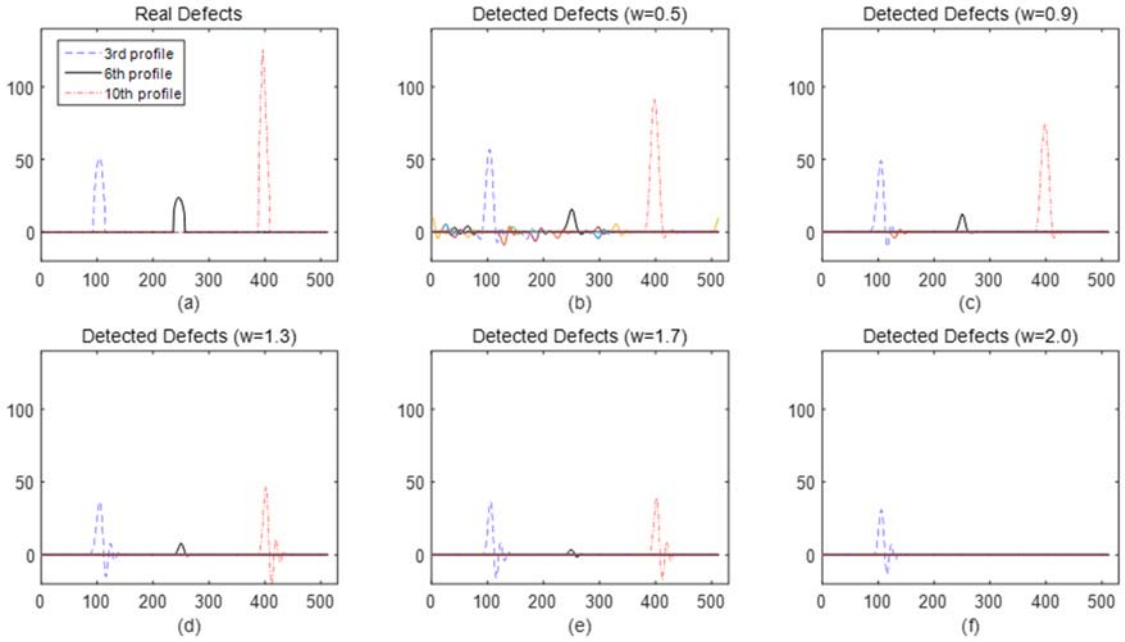
Figure 12. Comparison between the simulated noise and the estimated noise

Regarding noise, the simulated noise (underlying truth) and the estimated noise by the PMD are shown in Figure 12. We observe that the estimated noise after applying the PMD is quite consistent with the original simulated noise. We introduce an index to quantify the difference between the simulated noise and the estimated noise, $D_e = \|e_{\text{sim}} - e_{\text{est}}\|_F^2 / \|e_{\text{sim}}\|_F^2$, where e_{sim} and e_{est} are the simulated noise and the estimated noise, respectively, and $\|\cdot\|_F^2$ is square of the Frobenius norm. In the surrogated data analysis, $D_e = 0.3029$, which implies the non-uniform noise is modeled well in the PMD.

3.4.2 Sensitivity Analysis of the Bounds and the Penalized Parameter

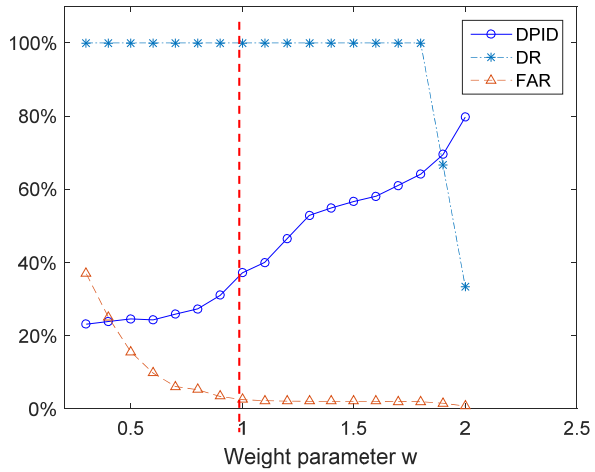
In order to evaluate detection accuracy of the PMD with different bounds choice, we fix the penalized parameter to 0.15. Then, we compare the simulated defects (underlying truth) introduced in the surrogated data and the detected defects in the decomposed defective effects under different bounds. Let bounds $\boldsymbol{\omega} = \boldsymbol{\omega}_1 = -\boldsymbol{\omega}_2 = w\mathbf{1}$,

where w is the weighting coefficient. The visual comparison can be found in Figure 13(a-f). The comparison based on the criteria, including detected peak intensity difference (DPID), detection rate (DR), and false alarm rate (FAR) are shown in Figure 14. Figure 13(a) corresponds to the simulated defects in the surrogated data. Figure 13(b-f) are associated with decomposed defective effects under $w = 0.5, 0.9, 1.3, 1.7, 2.0,$ respectively. From Figure 14, it is obvious that the DPID tends to become larger as the weighting coefficient w increases. The DR is lower than 100% when $w > 1.8$. The FAR becomes lower as the weighting coefficient w increases, and the FAR is close to zero when $w \geq 1$. The DR is high and meanwhile the FAR is low when $1 \leq w < 1.8$. Also the result is consistent with the detected defects shown visually in Figure 13.



(Note: all the x axis denotes Raman shift index, all the y axis denotes the Raman intensity (a.u.))

Figure 13. Comparison between real defects and detected defects in different profiles under different bounds

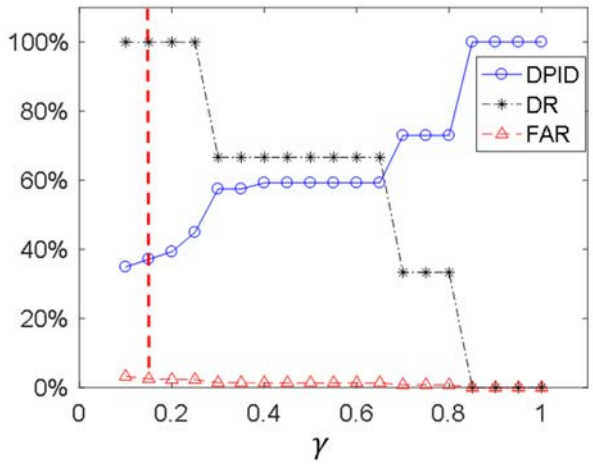


(Note: the red dashed vertical line denotes the selected weight parameter.)

Figure 14. Detected peak intensity difference, detection rate and false alarm rate at different weighting coefficient w

All the defects in different profiles can be detected accurately when $w = 1.3$ or 1.7. The detected Raman intensities tend to be a little smaller than the Raman intensities of the true defects. This can be explained from the following three aspects: (i) The most important one is due to the use of L_1 norm regularization, which leads to the soft-thresholding of the true defect information. (ii) The reduced amount of detected defects can also be explained as pure normal effects within the bounds. The detection sensitivity is controlled by the setting of bounds. The real defect information exceeds the bounds cannot be interpreted by the pure normal effects and it will be recognized by the defective effects. Thus the reduced amount of detected defects tend to be large when the bounds change from 0.5 to 2.0, which is consistent with the trend of detected peak intensity difference in Figure 14. If the weighting vector is $w = 0.5, 0.9$, there are false detection in other normal profiles shown in Figure 13(b,c), which will increase the type-I error during monitoring. While applying large weight vector such as $w = 2.0$ results in under-detection in the 6^{th} and the 10^{th} profiles shown in Figure 13(f), which will increase the type-II error during monitoring. A

reasonable setting of bounds can realize the trade-off between the detection power and the false alarm rate in the monitoring. Usually, we choose tighter bounds under the D-band, the G-band, or other important zones in a Raman spectra since they are closely related to the degree of functionalization and the degree of alignment for the quality of buckypaper. (iii) Although we choose oracle bounds for the PMD, the detected defects are smaller than the simulated defects. One reason of this is that the signal-dependent noise compensates part of the defective change in the profiles and the compensation results in the reduced amount of Raman intensity in defective effects.



(Note: the red dashed vertical line denotes the selected penalized parameter γ^*.)

Figure 15. Detected peak intensity difference, detection rate, and false alarm rate at different penalized parameters γ

We also explore the sensitivity analysis of the penalized parameter γ in the surrgoated analysis. In this case, we fix the bounds by $\boldsymbol{\omega}_1 = -\boldsymbol{\omega}_2 = \mathbf{1}$. The change of quantitative indices, including detected peak intensity difference (DPID), detection rate (DR), and false alarm rate (FAR), over the penalized parameter γ is shown in Figure 15. From Figure 15, it is obvious that the DPID tends to become larger as the penalized

coefficient γ increases, while the DR tends to become smaller. The reason of aforementioned properties is that larger penalized parameter results in more regularization on the L_1 norm of the anomalous coefficients. The pattern of the DR and the DPID have step/ladder patterns because there are three types of defects in different bands of the profiles. Once a defect is missed, the detection rate has a step decrease. For the false alarm rate, we can find that it decreases slightly as the increase of penalized parameter. That means the FAR is not highly dependent on the penalized parameter, and it is mainly influenced by the bounds ($\boldsymbol{B}_L$ and $\boldsymbol{B}_U$). Furthermore, the selected optimal penalized parameter $\gamma^* = 0.15$ obtained via the cross validation is shown in the red dashed line in Figure 15. The selection of γ^* can achieve very high detection rate with low false alarm rate and detected peak intensity difference.

3.5 Case Study

In this section, the proposed PMD is applied to a case study of in-line Raman spectra modeling and analysis. The setup of in-line Raman spectroscopy is shown in Figure 7. In the experimental set-up of the case study, Near Infra-Red (NIR) laser with a wavelength of 785nm and a power of 150mW were used to eliminate the effect of ambient light. A low magnification lens was used to achieve a larger focus tolerance.

Raman mapping technique is used to record Raman spectra in the measurement zone with a rectangular shape and a grid pattern. The measurement points and part of the quality information obtained by the PMD are shown in Figure 16. In Figure 16, ten samples with each sample containing ten measurement points are tested sequentially with a distance of 12 micrometers in the x-axis, which are marked with the red rectangular (dot line) or green

rectangular (line). All measurement locations are marked with a blue cross or a red circle. Raman spectrum with 512 Raman shifts (dimension) is collected for each measurement point. Within the same Raman sample, the distance between two measurement points is 1.2 micrometer in the y-axis, which is much smaller than the distance (12 micrometers) between samples in the x-axis. All the Raman spectra are collected based on a piece of single-wall CNTs buckypaper in the Renishaw confocal research Raman Microscopy with 785-nm laser source and 0.5-second exposure time for each measurement point.

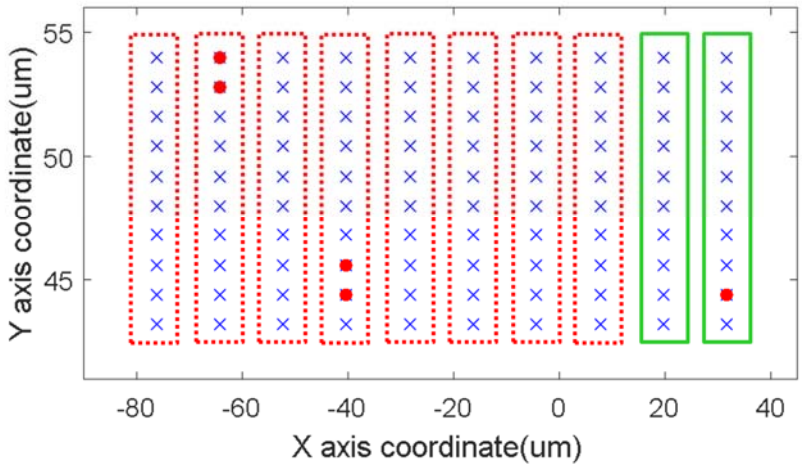


Figure 16. In-line Raman spectroscopy measurement zone and quality information obtained by the PMD

The PMD is implemented to decompose the in-line Raman spectra into fixed effects, normal effects, defective effects, and signal-dependent noise. The fixed effects of these ten Raman samples are shown in Figure 17. We can observe that there is a mean shift started from the 9th sample. The samples after the mean shift are marked by the green rectangular in Figure 16. The Raman intensity tends to be smaller than other normal samples, which likely comes from the measurement equipment (the focus depth change due to sample local

deformation) or product quality (the degree of functionalization, the degree of alignment changes, nanotube distribution and dispersion).

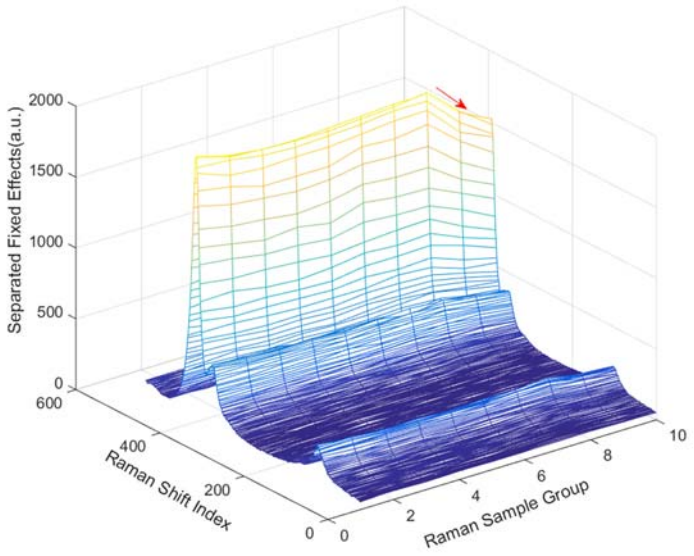


Figure 17. The fixed effects separated by the PMD in different Raman spectra

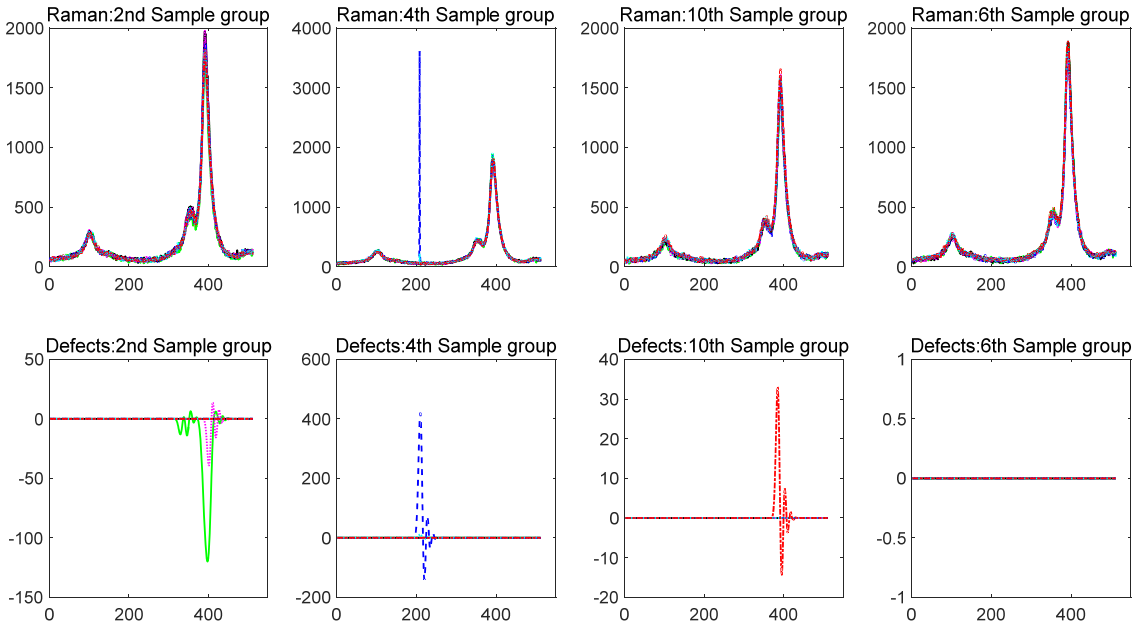


Figure 18. Original Raman spectra and defective effects separated by the PMD in the 2nd, 4th, 6th and 10th sample

The defective effects separated by the PMD show that the first two measurement points in the 2nd sample, the 8th and the 9th points in the 4th sample, and the 9th point in the 10th sample are defective points, which are marked as red circles in Figure 16. The original Raman spectra and defective effects by the PMD in the 2nd, the 4th, and the 10th samples are shown in Figure 18. The 6th sample, which is a normal Raman sample, is also shown in Figure 18 for the benchmark. We can observe that the defective effects remain constant for the 6th sample. We find that the defective effects occur in the G-band zone for the 2nd and 10th samples. The defects are due to impurity of raw material, different degree of functionalization, different alignment of carbon nanotubes, or bad nanotube dispersion. Moreover, a large spike is detected as defective effect between the D-band and the G-band for the 4th sample, which may result from some measurement errors, such as external light.

3.6 Conclusion

In the continuous CNTs buckypaper manufacturing process, in-line Raman spectroscopy is applied in collecting Raman spectra, which is a complicated profile data, to monitor the quality characterization and quantification of buckypaper. Three quality characteristics are primary concerns during the monitoring and control of buckypaper fabrication process: (i) quality consistency; (ii) within-sample uniformity; (iii) within-sample defect information.

The main contribution of this chapter is to propose a new PMD method to achieve multichannel profile detection and analysis based on in-line Raman spectroscopy in the continuous nanomanufacturing process. The proposed PMD can decompose the profiles into four parts: fixed effects, normal effects, defective effects, and signal-dependent noise.

Wavelet basis is applied as a design matrix in the penalized mixed model because of its ability for sparse representation of the Raman spectrum and its ability to handle non-smooth profiles. A weighted sum of the square loss function is applied to model the signal-dependent noise, and the L_1 norm regularization is implemented to induce the sparsity of the defective coefficients. The PMD combines the advantage of the LMM and the SSD, and performs better in multichannel profile detection based on the in-line Raman spectroscopy. An APG based algorithm is developed to efficiently handle the parameter estimation of the penalized mixed-effects decomposition with linear constraints. The developed algorithm is computationally efficient and meets the real time monitoring requirement during the CNTs buckypaper fabrication process. Furthermore, the sensitivity analysis of the bounds is conducted to demonstrate how to determine proper bounds to achieve a tradeoff between the detection power and the false alarm rate.

A surrogated data analysis is conducted to compare the performance of the PMD with the LMM and the SSD. The PMD decomposes the simulated profiles into four parts successfully. The suitable bounds can be determined by the sensitivity analysis and the product tolerance. We have used five criteria for comparison: (i) detection rate; (ii) false alarm rate; (iii) detected peak intensity difference; (iv) the mean square error; and (v) the computation time. The PMD can get a higher detection rate with a lower false alarm rate and a lower detected peak intensity difference compared to the benchmark methods. The PMD can realize 100% detection rate with 2.63% false alarm rate and 30.95% detected peak intensity difference. In addition, the mean square error and the computation time are lower than the benchmark methods. The case study shows that the decomposition based on the PMD is able to detect the mean shift of fixed effects and detect the defects in different

bands of different profiles. All these studies indicate that the PMD works well for the modeling and analysis of multichannel in-line Raman spectroscopy.

CHAPTER 4. TENSOR MIXED-EFFECTS MODEL AND APPLICATIONS IN NANOMANUFACTURING INSPECTION

Raman mapping technique has been used to perform in-line quality inspections of nanomanufacturing processes. In such an application, massive high-dimensional Raman mapping data with mixed effects is generated. In general, fixed effects and random effects in the multi-array Raman data are associated with different quality characteristics such as fabrication consistency, uniformity, defects, et al. The existing tensor decomposition methods cannot separate mixed effects, and existing mixed effects model can only handle matrix data but not high-dimensional multi-array data. In this chapter, we propose a tensor mixed effects (TME) model to analyze massive high-dimensional Raman mapping data with complex structure. The proposed TME model can (i) separate fixed effects and random effects in a tensor domain; (ii) explore the correlations along different dimensions; and (iii) realize efficient parameter estimation by a proposed iterative double Flip-Flop algorithm. We also investigate the properties of the TME model, existence and identifiability of parameter estimation. The numerical analysis demonstrates the efficiency and accuracy of the parameter estimation in the TME model. Convergence and asymptotic properties are discussed in the simulation and surrogate data analysis. The real case study shows an application of the TME model in quantifying the influence of alignment on carbon nanotubes buckypaper.

4.1 Introduction

Carbon nanotubes (CNTs) buckypaper is an important multifunctional platform material with great potential for creating lightweight and high-performance materials for various applications due to buckypaper's superior mechanical and electrical characteristics. One of the critical bottlenecks in the massive production and applications of high-quality buckypaper is quality inspection and monitoring of nanomanufacturing processes. The challenges include: (i) applying quick and accurate quality metrology to obtain information associated with microstructure, (ii) characterizing and analyzing in-line data to extract useful quality information for inspection and monitoring.

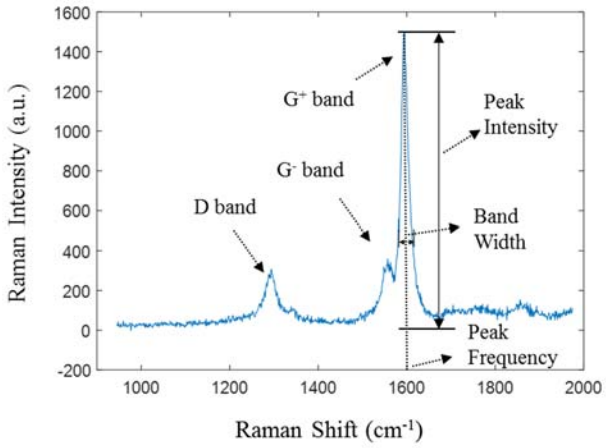


Figure 19. One Raman spectrum for Single-Walled CNTs buckypaper

As an effective characterization method for nanostructure information, Raman spectroscopy is very suitable for in-line quality inspection of nanomanufacturing processes. As an example, one Raman spectrum of single-walled CNTs buckypaper is shown in Figure 19. In the figure, the Raman peak intensity corresponds to material concentration and distribution; peak frequency is associated with molecular structure and phase; bandwidth is associated with crystallinity and phase (Salzer and Siesler 2009); intensity ratio of D-band and G-band can be affected by degree of functionalization (Cheng

et al. 2009). Therefore, numerous vital information about buckypaper quality is hidden in the Raman spectra data, which provides unprecedented opportunities for quality inspection, system informatics, and monitoring.

Due to the recent development of metrology technologies, Raman mapping (also called Raman spectral imaging) can be used to perform in-line quality inspection in continuous CNTs buckypaper nanomanufacturing processes. Raman mapping is a technique for generating detailed multi-array Raman spectra including numerous information about nanomaterials. Meanwhile, it is a challenging task to conduct data analytics, feature extraction, pattern recognition and in-line decision making, due to the high-dimensionality, large data size, as well as complex spatial and temporal correlation structures of Raman mapping. Specifically, in a Raman mapping measurement for single-walled CNTs buckypaper, about 600 Raman spectra can be collected per minute from a rectangular zone with a dimension of 10 micrometers by 60 micrometers. As shown in Figure 20, multiple measurement points are chosen from a rectangular zone, and each measurement point generates one Raman spectrum. Every Raman spectrum includes 1024 Raman shifts and intensities. The correlations along x/y directions of the rectangular zone are different due to the alignment of carbon nanotubes in the CNTs buckypaper. Meanwhile, the correlation along the Raman shift is different from the aforementioned spatial correlation.

According to the data structure of Raman mapping, tensor is an efficient mathematical tool for formulating the Raman mapping for nanomanufacturing inspection. Tensors (also called multidimensional arrays) have become increasingly important because they provide a concise mathematical framework for formulating the high-dimensional data.

Similar to the linear regression model in classical statistics, people use high-order tensor decompositions in high-dimensional statistics, such as CANDECOM/PARAFAC (CP) decomposition and Tucker decomposition, to separate different components inherent to the data. Kolda and Bader (2009) provided an overview of higher-order tensor decompositions, their applications, and available software. Corresponding to the generalized linear model (GLM) in statistics, Zhou et al. (2013) proposed a GLM model in the tensor domain, which extends the classical vector-valued covariate regression to an array-valued covariate regression. However, these tensor-based methods do not consider the multilevel variabilities (mixed effects) in the datasets.

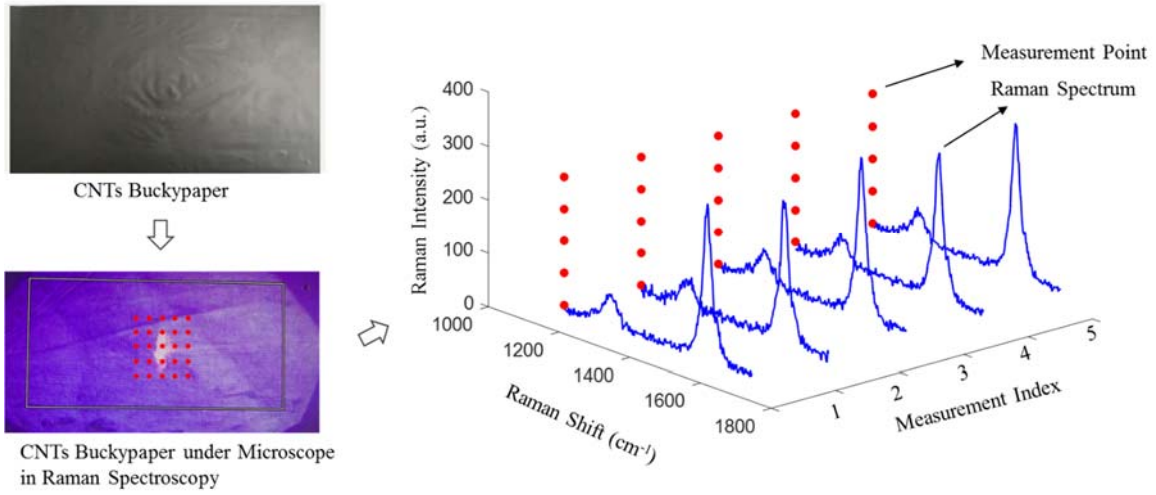


Figure 20. Raman mapping from the rectangular zone in CNTs buckypaper

A mixed effects model is a statistical model containing both fixed effects and random effects. It has nice properties, including (i) the capability to handle multilevel hierarchical data, such as longitudinal data with multiple measurements collected over time for an individual sensor; (ii) its ability to take complex association structures, including the correlation between different groups and correlation within an individual group, into

consideration. Thus, the mixed effects model is widely used in a variety of disciplines such as physics, biology, engineering and social sciences (Demidenko 2013; Galecki and Burzykowski 2013). However, the classical mixed effects model treats multivariate data as a vector or a matrix, which is insufficient for analysis of high-dimensional data, such as tensor-type Raman mapping with its high dimensionality and complex correlations. Thus we need to develop a novel tensor mixed effects (TME) model that can explore mixed effects in the tensor domain.

We emphasize the motivation of developing a TME model with an example in nanomanufacturing. Raman mapping data are collected to inspect the quality of continuously fabricated CNTs buckypaper. There are multiple components in the data that are associated with different critical quality characteristics. Specifically, fixed effects measure the fabrication consistency of quality features (such as degree of alignment, degree of functionalization, nanotube distribution, and dispersion). This indicates whether there is a gradual mean shift in the roll-to-roll fabrication process of CNTs buckypaper. In addition, random effects are relevant to the uniformity and defect information. The uniformity pertains to the status of the quality indices, while the defect information consists of the number and the pattern of defects in the CNTs buckypaper (Yue et al. 2018). Therefore, it is necessary to use a mixed effects model to decompose different effects in the Raman mapping data. From another point of view, Raman mapping data have tensor structures. One Raman mapping dataset usually contains multiple dimensions: two measurement coordinates, Raman shift (frequency) and Raman intensity. If matricization or vectorization is conducted to process the Raman data, a classical mixed effects model can be developed. However, this vectorized linear mixed effects (vLME) model has three limitations: (i) the

dimension after matricization/vectorization becomes very high and a large sample size is required for accurate parameter estimation; (ii) the computation cost will be large, and it cannot meet the needs of in-line inspection and monitoring; (iii) the transformation alters the inherent multi-way correlation structures, which makes the correlation along different dimensions unobtainable. To overcome these three limitations, this chapter proposes the tensor mixed effects (TME) model.

The TME model can effectively and efficiently explore multilevel variabilities (including fixed effects and random effects) inherent to tensor-structured high-dimensional data. It can be regarded as a logical extension from a vector-valued or matrix-valued mixed effects model to an array-valued mixed effects model. It is a challenging task to develop the TME model because (i) it deals with high-dimensional datasets with tensor structure; (ii) an efficient algorithm is required to do parameter estimation; (iii) it is necessary to ensure the identifiability of multi-dimensional correlations. In this chapter, we propose the TME model and explore its properties. An iterative Flip-Flop algorithm is proposed for parameter estimation.

The remainder of this chapter is organized as follows. Section 4.2 introduces basic tensor notation and preliminaries, and further proposes the TME model. Section 4.3 describes the proposed maximum likelihood estimation (MLE) algorithm for the TME model, in addition to investigating the existence of the MLE and the identifiability of the TME model. In Section 4.4, a double Flip-Flop algorithm is proposed to conduct parameter estimation of the TME model. In addition, initialization and convergence criteria of the algorithm are provided. Sections 4.5 presents a numerical simulation, a surrogated data

analysis and a real case study of Raman mapping to test the performance of the TME model. Finally, a brief summary is provided in Section 4.6.

4.2 Tensor Mixed-effects Model

In this section, we first introduce the tensor notation and preliminaries. Then, we propose the TME model and explore the random distribution of tensor responses. Next, we discuss the maximum likelihood estimation (MLE) for the TME model, the conditions for the existence of the MLE, and the constraints to ensure the identifiability.

4.2.1 Tensor Notation and Preliminaries

In this section, basic notations, definitions, and matrix/array operators in multilinear (tensor) algebra are introduced and summarized. The terminology used here remains as consistent as possible with the terminology of previous publications (Kolda and Bader 2009; Zhou et al. 2013) in the area of tensor algebra. Scalars are denoted by lowercase italic letters, e.g., a; vectors by lowercase italic boldface letters, e.g., $\boldsymbol{a}$; matrices by uppercase italic boldface letters, e.g., $\boldsymbol{A}$; and tensors by calligraphic letters, e.g., $\boldsymbol{\mathcal{X}}$. The order of a tensor is the number of dimensions (modes). For example, an K-order tensor is denoted by $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$, where I_k denotes the k-mode dimension of $\boldsymbol{\mathcal{X}}$. The ith entry of a vector $\boldsymbol{a}$ is denoted by a_i, the element (i, j) of a matrix $\boldsymbol{A}$ is denoted by a_{ij}, and the element (i, j, k) of a third-order tensor $\boldsymbol{\mathcal{X}}$ is denoted by x_{ijk}. Indices range from 1 to their capital versions, e.g., $i = 1, \cdots, I$.

Matricization, also known as unfolding or flattening, is the process of reordering the elements of a tensor into a matrix (Kolda and Bader 2009). The k-mode matricization

of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ is denoted by $\boldsymbol{X}_{(k)}$. $\text{vec}(\boldsymbol{\mathcal{X}})$ is the vectorization of a tensor $\boldsymbol{\mathcal{X}}$. The k-mode product of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_K}$ with a matrix $\boldsymbol{U} \in \mathbb{R}^{J \times I_k}$ is denoted by $\boldsymbol{\mathcal{X}} \times_k \boldsymbol{U}$ and elementwise, we have $(\boldsymbol{\mathcal{X}} \times_k \boldsymbol{U})_{i_1 \cdots i_{k-1} j i_{k+1} \cdots i_K} = \sum_{i_k=1}^{I_k} x_{i_1 \cdots i_k \cdots i_K} u_{j i_k}$, where all the indices range from 1 to their capital versions, e.g., the index j goes from $1,2,\ldots,J$, and the index i_k goes from $1,2,\ldots,I_k$. The kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are denoted by $\boldsymbol{A} \otimes \boldsymbol{B}$. The kronecker product is an operation on two matrices resulting in a block matrix and it is a generalization of the outer product.

4.2.2 Tensor Mixed-effects Model

Firstly, we consider a TME model for the third-order tensor data

$$\boldsymbol{\mathcal{Y}}_i = \boldsymbol{\mathcal{F}} \times_1 \boldsymbol{A}_i^{(1)} \times_2 \boldsymbol{A}_i^{(2)} \times_3 \boldsymbol{A}_i^{(3)} + \boldsymbol{\mathcal{R}}_i \times_1 \boldsymbol{B}_i^{(1)} \times_2 \boldsymbol{B}_i^{(2)} \times_3 \boldsymbol{B}_i^{(3)} + \boldsymbol{\mathcal{E}}_i \tag{24}$$

where the ith response tensor is $\boldsymbol{\mathcal{Y}}_i \in \mathbb{R}^{J \times K \times L}$ with $i = 1,\cdots,N$; N is the sample size; the fixed effects core tensor is $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{P_1 \times Q_1 \times R_1}$; $\boldsymbol{A}_i^{(1)} \in \mathbb{R}^{J \times P_1}$, $\boldsymbol{A}_i^{(2)} \in \mathbb{R}^{K \times Q_1}$, $\boldsymbol{A}_i^{(3)} \in \mathbb{R}^{L \times R_1}$ are the design (factor) matrices for the fixed effects; the random effects core tensor is denoted by $\boldsymbol{\mathcal{R}}_i \in \mathbb{R}^{P_2 \times Q_2 \times R_2}$, and the corresponding design (factor) matrices for the random effects by $\boldsymbol{B}_i^{(1)} \in \mathbb{R}^{J \times P_2}$, $\boldsymbol{B}_i^{(2)} \in \mathbb{R}^{K \times Q_2}$, $\boldsymbol{B}_i^{(3)} \in \mathbb{R}^{L \times R_2}$; the tensor for the residual errors is denoted by $\boldsymbol{\mathcal{E}}_i \in \mathbb{R}^{J \times K \times L}$. Both the fixed effects and the random effects can be regarded as Tucker decompositions of original fixed/random effects. We also denote the Tucker decomposition by $\left[\!\left[\boldsymbol{\mathcal{F}}; \boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}\right]\!\right] = \boldsymbol{\mathcal{F}} \times_1 \boldsymbol{A}_i^{(1)} \times_2 \boldsymbol{A}_i^{(2)} \times_3 \boldsymbol{A}_i^{(3)}$. Same to the requirement in Tucker decomposition, both the design matrices $\boldsymbol{A}_i^{(\mathrm{j})}, j = 1,2,3$ and $\boldsymbol{B}_i^{(\mathrm{j})}, j = 1,2,3$ are chosen to be orthogonal. Usually, the higher order Tucker

decomposition follows similar structure as the third-order decomposition. Thus, it is straightforward to extend the third-order TME model to higher dimensional analysis.

Similar to the classical mixed effects model, we assume that the specification of the random effects core tensor $\boldsymbol{\mathcal{R}}_i$ and the residual errors tensor $\boldsymbol{\mathcal{E}}_i$ follow tensor normal distributions. Particularly, the tensor normal distribution of random effects core tensor $\boldsymbol{\mathcal{R}}_i$ is $\boldsymbol{N}_{P_2,Q_2,R_2}(\boldsymbol{\mathcal{O}}; \boldsymbol{\Sigma}_{\mathrm{r}}, \boldsymbol{\Psi}_{\mathrm{r}}, \boldsymbol{\Omega}_{\mathrm{r}})$, where the mean tensor $\boldsymbol{\mathcal{O}}$ is a zero tensor, and the covariance matrices along different dimensions $\boldsymbol{\Sigma}_{\mathrm{r}} \in \mathbb{R}^{P_2 \times P_2}, \boldsymbol{\Psi}_{\mathrm{r}} \in \mathbb{R}^{Q_2 \times Q_2}, \boldsymbol{\Omega}_{\mathrm{r}} \in \mathbb{R}^{R_2 \times R_2}$ are positive definite. We know, from the properties of tensor normal distribution, $\mathrm{vec}(\boldsymbol{\mathcal{R}}_i)$ is distributed as a multivariate normal distribution with mean $\mathrm{vec}(\boldsymbol{\mathcal{O}})$ and covariance matrix $\boldsymbol{\Omega}_{\mathrm{r}} \otimes \boldsymbol{\Psi}_{\mathrm{r}} \otimes \boldsymbol{\Sigma}_{\mathrm{r}}$. We write that

$$\boldsymbol{\mathcal{R}}_i \backsim \boldsymbol{N}_{P_2,Q_2,R_2}(\boldsymbol{\mathcal{O}}; \boldsymbol{\Sigma}_{\mathrm{r}}, \boldsymbol{\Psi}_{\mathrm{r}}, \boldsymbol{\Omega}_{\mathrm{r}}) \text{ if } \mathrm{vec}(\boldsymbol{\mathcal{R}}_i) \backsim \boldsymbol{N}_{P_2 Q_2 R_2}(\mathrm{vec}(\boldsymbol{\mathcal{O}}), \boldsymbol{\Omega}_{\mathrm{r}} \otimes \boldsymbol{\Psi}_{\mathrm{r}} \otimes \boldsymbol{\Sigma}_{\mathrm{r}}).$$

Similarly, the distribution of the residual errors tensor $\boldsymbol{\mathcal{E}}_i$ is $\boldsymbol{N}_{J,K,L}(\boldsymbol{\mathcal{O}}; \boldsymbol{\Sigma}_{\varepsilon}, \boldsymbol{\Psi}_{\varepsilon}, \boldsymbol{\Omega}_{\varepsilon})$, and the noise covariance matrices along different dimensions are $\boldsymbol{\Sigma}_{\varepsilon} \in \mathbb{R}^{J \times J}, \boldsymbol{\Psi}_{\varepsilon} \in \mathbb{R}^{K \times K}, \boldsymbol{\Omega}_{\varepsilon} \in \mathbb{R}^{L \times L}$. Thus, $\mathrm{vec}(\boldsymbol{\mathcal{E}}_i) \backsim \boldsymbol{N}_{JKL}(\mathrm{vec}(\boldsymbol{\mathcal{O}}), \boldsymbol{\Omega}_{\varepsilon} \otimes \boldsymbol{\Psi}_{\varepsilon} \otimes \boldsymbol{\Sigma}_{\varepsilon})$. Moreover, we assume that the random effects core tensor and residual errors tensor are independent of each other. According to the descriptions above, we can find that the parameter size of the TME model in Equation (24) is $P_1 \times Q_1 \times R_1 + (P_2 + P_2^2 + Q_2 + Q_2^2 + R_2 + R_2^2)/2 + J + K + L$, with the assumption that the covariance matrices $\boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\Psi}_{\varepsilon}$, $\boldsymbol{\Omega}_{\varepsilon}$ are diagonal. While the parameter size of the corresponding vectorized linear mixed effects (vLME) model is $P_1 \times Q_1 \times R_1 + (P_2 Q_2 R_2 + P_2^2 Q_2^2 R_2^2)/2 + JKL$, with the assumption that the covariance matrix of noise term is diagonal. Therefore, the parameter size of the vLME model is larger than the

parameter size of the TME model. A larger sample size will be needed for the vLME model to realize accurate parameter estimation.

In addition to the fixed effects core tensor and design matrices, the TME model includes two sources of random components: the random effects accounting for covariance along different dimensions, and the residual errors $\boldsymbol{\mathcal{E}}_i$ relevant to the inevitable random noise. Based on the properties of tensor normal distribution, we can derive the random distribution of $\boldsymbol{\mathcal{Y}}_i$, as shown in Proposition 4.

Proposition 4. The response tensor in Equation (24) follows a tensor normal distribution, that is

$$\begin{aligned}\boldsymbol{\mathcal{Y}}_i \backsim \boldsymbol{N}_{J,K,L}\Big(\Big[\!\Big[\boldsymbol{\mathcal{F}}; \boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}\Big]\!\Big]; \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T} + \boldsymbol{\Sigma}_{\varepsilon}, \boldsymbol{B}_i^{(2)}\boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T} \\ + \boldsymbol{\Psi}_{\varepsilon}, \boldsymbol{B}_i^{(3)}\boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T} + \boldsymbol{\Omega}_{\varepsilon}\Big)\end{aligned} \tag{25}$$

Proof: please see the appendix A.4.

For simplification, we define $\widetilde{\boldsymbol{\mathcal{F}}} = \Big[\!\Big[\boldsymbol{\mathcal{F}}; \boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}\Big]\!\Big]$, $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T} + \boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\Psi}_i = \boldsymbol{B}_i^{(2)}\boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T} + \boldsymbol{\Psi}_{\varepsilon}$, and $\boldsymbol{\Omega}_i = \boldsymbol{B}_i^{(3)}\boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T} + \boldsymbol{\Omega}_{\varepsilon}$. The total covariance matrices $\boldsymbol{\Sigma}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i$ are positive definite. Thus, the response tensor distribution in Equation (25) can be written as $\boldsymbol{\mathcal{Y}}_i \backsim \boldsymbol{N}_{J,K,L}\big(\widetilde{\boldsymbol{\mathcal{F}}}; \boldsymbol{\Sigma}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i\big)$. It can be further described in matrix form using three different modes shown as

$$\boldsymbol{Y}_{i(1)} \backsim \boldsymbol{N}_{J,KL}\Big(\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)})^T; \boldsymbol{\Sigma}_i, \boldsymbol{\Omega}_i \otimes \boldsymbol{\Psi}_i\Big) \tag{26}$$

$$\boldsymbol{Y}_{i(2)} \backsim \boldsymbol{N}_{K,JL}\Big(\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(1)})^T; \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i \otimes \boldsymbol{\Sigma}_i\Big) \tag{27}$$

$$\boldsymbol{Y}_{i(3)} \backsim \boldsymbol{N}_{L,JK}\Big(\boldsymbol{A}_i^{(3)}\boldsymbol{F}_{(3)}(\boldsymbol{A}_i^{(2)} \otimes \boldsymbol{A}_i^{(1)})^T; \boldsymbol{\Omega}_i, \boldsymbol{\Psi}_i \otimes \boldsymbol{\Sigma}_i\Big) \tag{28}$$

where $\boldsymbol{Y}_{i(k)}$ and $\boldsymbol{F}_{(k)}$ (k=1,2,3.) are the k-mode matricization of the tensor $\boldsymbol{\mathcal{Y}}_i$ and $\boldsymbol{\mathcal{F}}$. Obviously, Equations (26-28) show $\boldsymbol{Y}_{i(1)}$, $\boldsymbol{Y}_{i(2)}$, $\boldsymbol{Y}_{i(3)}$ follow matrix normal distributions.

In this section, we proposed the TME model and specified the distribution of random effects core tensor and errors tensor. We also derived the random distribution of the response tensors in Proposition 4, which lays a foundation for inference in Section 4.3.

4.3 Inference of the TME Model

This section discusses how to estimate the parameters in the TME model by using the maximum likelihood estimation (MLE). Generally speaking, the parameter estimation of a TME model involves three steps: (i) constructing a log-likelihood function for the MLE with relevant probability distribution functions; (ii) deriving the MLE of fixed effects and total covariance matrices along different dimensions; and (iii) obtaining the MLE for covariance matrices of residual errors based on the conditional probability distribution.

4.3.1 MLE of Fixed Effects and Total Covariance Matrices

We know that the response tensor $\boldsymbol{\mathcal{Y}}_i$ follows the tensor normal distribution in Equation (25) and the k-mode matricization $\boldsymbol{Y}_{i(k)}$ follows the matrix normal distributions as shown in Equations (26-28). Thus, we have Proposition 5.

Proposition 5. The log-likelihood functions of Equations (26-28) are the same, and can be represented as

$$\boldsymbol{l}_i = -\frac{JKL}{2}\log 2\pi - \frac{JK}{2}\log|\boldsymbol{\Omega}_i| - \frac{JL}{2}\log|\boldsymbol{\Psi}_i| - \frac{KL}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\Bigg(\mathrm{vec}\Big(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\Big(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\Big)^T\Big)\Bigg)^T (\boldsymbol{\Omega}_i^{-1}\otimes\boldsymbol{\Psi}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathrm{vec}\Big(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\Big(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\Big)^T\Big). \tag{29}$$

Proof: please see the appendix A.5.

Maximization of the log-likelihood function (29) yields the MLE estimators, which are shown in Proposition 6.

Proposition 6. Given the response tensors $\boldsymbol{\mathcal{Y}}_i$, and the basis $\boldsymbol{A}_i^{(1)}$, $\boldsymbol{A}_i^{(2)}$, $\boldsymbol{A}_i^{(3)}$ with $i = 1,\cdots,N$, the maximum likelihood estimator of $\mathrm{vec}(\boldsymbol{\mathcal{F}})$ is

$$\begin{aligned}\mathrm{vec}(\widehat{\boldsymbol{\mathcal{F}}}) = &\left(\sum_{i=1}^{N}\left(\boldsymbol{A}_i^{(3)^T}\boldsymbol{\Omega}_i^{-1}\boldsymbol{A}_i^{(3)}\right)\otimes\left(\boldsymbol{A}_i^{(2)^T}\boldsymbol{\Psi}_i^{-1}\boldsymbol{A}_i^{(2)}\right)\otimes\left(\boldsymbol{A}_i^{(1)^T}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{A}_i^{(1)}\right)\right)^{-1} \\ &\cdot\left(\sum_{i=1}^{N}\left(\boldsymbol{A}_i^{(3)^T}\boldsymbol{\Omega}_i^{-1}\right)\otimes\left(\boldsymbol{A}_i^{(2)^T}\boldsymbol{\Psi}_i^{-1}\right)\otimes\left(\boldsymbol{A}_i^{(1)^T}\boldsymbol{\Sigma}_i^{-1}\right)\cdot\mathrm{vec}(\boldsymbol{\mathcal{Y}}_i)\right)\end{aligned} \tag{30}$$

When $\boldsymbol{B}_i^{(1)}$, $\boldsymbol{B}_i^{(2)}$, and $\boldsymbol{B}_i^{(3)}$ are constant for all $i = 1,\cdots,N$, and setting $\boldsymbol{B}_i^{(1)} = \boldsymbol{B}^{(1)}$, $\boldsymbol{B}_i^{(2)} = \boldsymbol{B}^{(2)}$, $\boldsymbol{B}_i^{(3)} = \boldsymbol{B}^{(3)}$ for $i = 1,\cdots,N$. For simplification, we define $\widehat{\widehat{\boldsymbol{\mathcal{F}}}} = \left[\!\left[\widehat{\boldsymbol{\mathcal{F}}}; \boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}\right]\!\right]$. The maximum likelihood estimators of $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, and $\boldsymbol{\Omega}_i$ are

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{KLN}\sum_{i=1}^{N}\left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(1)}\cdot\left(\widehat{\boldsymbol{\Omega}}_i^{-1}\otimes\widehat{\boldsymbol{\Psi}}_i^{-1}\right)\cdot\left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(1)}^T \tag{31}$$

$$\widehat{\boldsymbol{\Psi}}_i = \frac{1}{JLN}\sum_{i=1}^{N}\left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(2)}\cdot\left(\widehat{\boldsymbol{\Omega}}_i^{-1}\otimes\widehat{\boldsymbol{\Sigma}}_i^{-1}\right)\cdot\left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(2)}^T \tag{32}$$

$$\widehat{\boldsymbol{\Omega}}_i = \frac{1}{JKN}\sum_{i=1}^{N}\left(\boldsymbol{\mathcal{Y}}_i - \widehat{\bar{\boldsymbol{\mathcal{F}}}}\right)_{(3)} \cdot \left(\widehat{\boldsymbol{\Psi}}_i^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_i^{-1}\right) \cdot \left(\boldsymbol{\mathcal{Y}}_i - \widehat{\bar{\boldsymbol{\mathcal{F}}}}\right)_{(3)}^{T} \tag{33}$$

If both $\boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}$ and $\boldsymbol{B}_i^{(1)}, \boldsymbol{B}_i^{(2)}, \boldsymbol{B}_i^{(3)}$ are constant for all $i = 1,\cdots,N$, then $\bar{\boldsymbol{\mathcal{Y}}}$ is the mean response tensor, and the maximum likelihood estimators of $\boldsymbol{\Sigma}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i$ are

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{KLN}\sum_{i=1}^{N}(\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(1)} \cdot \left(\widehat{\boldsymbol{\Omega}}_i^{-1} \otimes \widehat{\boldsymbol{\Psi}}_i^{-1}\right) \cdot (\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(1)}^{T} \tag{34}$$

$$\widehat{\boldsymbol{\Psi}}_i = \frac{1}{JLN}\sum_{i=1}^{N}(\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(2)} \cdot \left(\widehat{\boldsymbol{\Omega}}_i^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_i^{-1}\right) \cdot (\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(2)}^{T} \tag{35}$$

$$\widehat{\boldsymbol{\Omega}}_i = \frac{1}{JKN}\sum_{i=1}^{N}(\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(3)} \cdot \left(\widehat{\boldsymbol{\Psi}}_i^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_i^{-1}\right) \cdot (\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(3)}^{T} \tag{36}$$

Proof: please see the appendix A.6.

Moreover, we can show that the estimator $\text{vec}(\widehat{\boldsymbol{\mathcal{F}}})$ given in Equation (30) is uniquely determined regardless of the parametrization of the covariance matrices, which is explored in appendix A.6.

From Equations (31-36), we can see that the estimators of covariance matrices are cross-related. A Flip-Flop type algorithm is designed to compute them. We will describe the algorithm in Section 4.4. Before that, we continue to explore the MLE for the covariance matrices of residual errors based on conditional probability distribution.

4.3.2 MLE of Random Effects and Residual Covariance Matrices

After finishing the estimation of the fixed effects and total covariance matrices, we consider the estimation for the covariance matrices of random effects and residual errors. The distribution of random effects core tensor $\boldsymbol{\mathcal{R}}_i$ conditioned on response tensors $\boldsymbol{\mathcal{Y}}_i$ $(i = 1,\cdots,N)$ follows a tensor normal distribution. Assuming $\boldsymbol{\mathcal{Y}}_i$ and $\widetilde{\boldsymbol{\mathcal{F}}}$ are known, the estimation of $\boldsymbol{\mathcal{R}}_i$ is the expectation of $\boldsymbol{\mathcal{R}}_i|\boldsymbol{\mathcal{Y}}_i$ and it can obtain

$$\widehat{\boldsymbol{\mathcal{R}}}_i = \left[\!\left[\boldsymbol{\mathcal{Y}}_i - \widetilde{\boldsymbol{\mathcal{F}}}; \boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T}\boldsymbol{\Sigma}_i^{-1}, \boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T}\boldsymbol{\Psi}_i^{-1}, \boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T}\boldsymbol{\Omega}_i^{-1}\right]\!\right]. \tag{37}$$

The distribution of $\boldsymbol{\mathcal{Y}}_i - \widetilde{\boldsymbol{\mathcal{F}}} - \left[\!\left[\boldsymbol{\mathcal{R}}_i; \boldsymbol{B}_i^{(1)}, \boldsymbol{B}_i^{(2)}, \boldsymbol{B}_i^{(3)}\right]\!\right]$ conditioned on the random effects core tensor $\boldsymbol{\mathcal{R}}_i$ is a tensor normal distribution given by

$$\left(\boldsymbol{\mathcal{Y}}_i - \widetilde{\boldsymbol{\mathcal{F}}} - \left[\!\left[\boldsymbol{\mathcal{R}}_i; \boldsymbol{B}_i^{(1)}, \boldsymbol{B}_i^{(2)}, \boldsymbol{B}_i^{(3)}\right]\!\right]\right)|\boldsymbol{\mathcal{R}}_i \backsim \boldsymbol{N}_{J,K,L}(\boldsymbol{\mathcal{O}}; \boldsymbol{\Sigma}_{\varepsilon}, \boldsymbol{\Psi}_{\varepsilon}, \boldsymbol{\Omega}_{\varepsilon}).$$

For simplification, we define $\widetilde{\boldsymbol{\mathcal{R}}}_i = \left[\!\left[\boldsymbol{\mathcal{R}}_i; \boldsymbol{B}_i^{(1)}, \boldsymbol{B}_i^{(2)}, \boldsymbol{B}_i^{(3)}\right]\!\right]$, $\widehat{\widetilde{\boldsymbol{\mathcal{R}}}}_i = \left[\!\left[\widehat{\boldsymbol{\mathcal{R}}}_i; \boldsymbol{B}_i^{(1)}, \boldsymbol{B}_i^{(2)}, \boldsymbol{B}_i^{(3)}\right]\!\right]$. Similar to Proposition 6, we have the maximum likelihood estimators of $\boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\Psi}_{\varepsilon}$, and $\boldsymbol{\Omega}_{\varepsilon}$ are

$$\widehat{\boldsymbol{\Sigma}}_{\varepsilon} = \frac{1}{KLN}\sum_{i=1}^{N}\left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widetilde{\boldsymbol{\mathcal{F}}}} - \widehat{\widetilde{\boldsymbol{\mathcal{R}}}}\right)_{(1)} \cdot \left(\widehat{\boldsymbol{\Omega}}_{\varepsilon}^{-1} \otimes \widehat{\boldsymbol{\Psi}}_{\varepsilon}^{-1}\right) \cdot \left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widetilde{\boldsymbol{\mathcal{F}}}} - \widehat{\widetilde{\boldsymbol{\mathcal{R}}}}\right)_{(1)}^{T} \tag{38}$$

$$\widehat{\boldsymbol{\Psi}}_{\varepsilon} = \frac{1}{JLN}\sum_{i=1}^{N}\left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widetilde{\boldsymbol{\mathcal{F}}}} - \widehat{\widetilde{\boldsymbol{\mathcal{R}}}}\right)_{(2)} \cdot \left(\widehat{\boldsymbol{\Omega}}_{\varepsilon}^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{-1}\right) \cdot \left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widetilde{\boldsymbol{\mathcal{F}}}} - \widehat{\widetilde{\boldsymbol{\mathcal{R}}}}\right)_{(2)}^{T} \tag{39}$$

$$\widehat{\boldsymbol{\Omega}}_{\varepsilon} = \frac{1}{JKN}\sum_{i=1}^{N}\left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widetilde{\boldsymbol{\mathcal{F}}}} - \widehat{\widetilde{\boldsymbol{\mathcal{R}}}}\right)_{(3)} \cdot \left(\widehat{\boldsymbol{\Psi}}_{\varepsilon}^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{-1}\right) \cdot \left(\boldsymbol{\mathcal{Y}}_i - \widehat{\widetilde{\boldsymbol{\mathcal{F}}}} - \widehat{\widetilde{\boldsymbol{\mathcal{R}}}}\right)_{(3)}^{T} \tag{40}$$

By comparing Equations (38-40) with Equations (31-33), we notice similar patterns. The mean components change from $\boldsymbol{\mathcal{Y}}_i - \widehat{\widetilde{\boldsymbol{\mathcal{F}}}}$ to $\boldsymbol{\mathcal{Y}}_i - \widehat{\widetilde{\boldsymbol{\mathcal{F}}}} - \widehat{\widetilde{\boldsymbol{\mathcal{R}}}}$ based on the

estimation of random effects. After that, a progressive estimation for covariance matrices $\boldsymbol{\Sigma}_{\varepsilon}, \boldsymbol{\Psi}_{\varepsilon}, \boldsymbol{\Omega}_{\varepsilon}$ are obtained.

We know that the covariance matrices $(\boldsymbol{\Sigma}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i)$ and $(\boldsymbol{\Sigma}_{\varepsilon}, \boldsymbol{\Psi}_{\varepsilon}, \boldsymbol{\Omega}_{\varepsilon})$ should be positive definite. In order to ensure the positive definite property in Equations (31-33,38-40), the existence of the MLE should be explored. This is shown in Section 4.3.2. Based on $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T} + \boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\Psi}_i = \boldsymbol{B}_i^{(2)}\boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T} + \boldsymbol{\Psi}_{\varepsilon}$, $\boldsymbol{\Omega}_i = \boldsymbol{B}_i^{(3)}\boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T} + \boldsymbol{\Omega}_{\varepsilon}$, we know that the covariance matrices of random effects and residual errors are not unique. The identifiability should be investigated, which is discussed in Section 4.3.3.

4.3.3 Existence of the MLE

Finding the estimation of the average component and fixed effects core tensor $\widehat{\boldsymbol{\mathcal{F}}}$, are straightforward given the positive definite covariance matrices. Hereafter, we focus on the exploration of the existence of MLE for the total covariance matrices $(\widehat{\boldsymbol{\Sigma}}_i, \widehat{\boldsymbol{\Psi}}_i, \widehat{\boldsymbol{\Omega}}_i)$, shown in Equations (31-36). A necessary condition for the existence of the MLE can be derived based on the paper (Manceur and Dutilleul 2013), which is demonstrated in Proposition 7.

Proposition 7. If maximum likelihood estimators for the covariance matrices $\boldsymbol{\Sigma}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i$ in the TME model (24) exist, the sample size N of the response tensors $\boldsymbol{\mathcal{Y}}_i$ $(i = 1,\cdots,N)$ satisfies the condition

$$N \geq \max\left(\frac{J}{KL}, \frac{K}{JL}, \frac{L}{JK}\right) + 1.$$

The Proof is straightforward according to the conclusion in the paper (Manceur and Dutilleul 2013). Although the condition shown in Proposition 7 is necessary for the existence of the MLE, it is not sufficient because that it cannot ensure that all the iterations of the algorithm have full rank matrices. Similar to the existence of the MLE for the model with Kronecker product covariance structure (Roś et al. 2016), it could happen that covariance matrices in the updated iterations do not have a full rank with the likelihood of the TME model converging to the supremum. The reason is that the space $\{\boldsymbol{\Omega}_i \otimes \boldsymbol{\Psi}_i \otimes \boldsymbol{\Sigma}_i : \boldsymbol{\Sigma}_i \in \mathbb{R}^{J\times J}, \boldsymbol{\Psi}_i \in \mathbb{R}^{K\times K}, \boldsymbol{\Omega}_i \in \mathbb{R}^{L\times L}; \boldsymbol{\Omega}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Sigma}_i \text{ are positive definite}\}$ with any norm is not closed. If we choose a stronger condition, for a space $\mathbb{K}$ (equipped with the Frobenius norm) of positive definite $JKL \times JKL$ matrices that have a kronecker structure such that $\boldsymbol{\Omega}_i \otimes \boldsymbol{\Psi}_i \otimes \boldsymbol{\Sigma}_i \in \mathbb{K}$, where $\boldsymbol{\Omega}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Sigma}_i$ are also positive definite, Then $\mathbb{K}$ is closed, according to the natural extension of (Roś et al. 2016). Based on this conclusion, we can formulate the sufficient condition for the existence of the MLE for the total covariance matrices, as shown in Proposition 8.

Proposition 8. The response tensor $\boldsymbol{\mathcal{Y}}_i$ $(i = 1,\cdots,N)$ satisfies the model in Equation (24). If $N \geq JKL$, maximum likelihood estimators for the covariance matrices $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Omega}_i$ in the TME model exist with probability 1.

The proof is straightforward by using the conclusion (Burg et al. 1982), and it is an extension of Theorem 3 in Page 6 of Roś et al. (2016).

In summary of Propositions 7 and 8, if $N < \max\left(\frac{J}{KL}, \frac{K}{JL}, \frac{L}{JK}\right) + 1$, the MLE of covariance matrices does not exist. However, if $N \geq JKL$, the MLE exists with probability 1. Moreover, the dimensions of tensor samples are usually large, and $N \geq JKL$ is hard to

guarantee in practice. When the covariance matrices satisfy special structures, such as diagonal structures, the sufficient condition of existence will be changed. We have Proposition 9 to illustrate the existence conditions with the additional assumptions of diagonal matrices.

Proposition 9. The response tensors $\boldsymbol{\mathcal{Y}}_i$ $(i = 1,\cdots,N)$ satisfy the model in Equation (24) with the additional assumption that $\boldsymbol{\Sigma}_i$ is diagonal. If $N \geq \max\left(KL, \max\left(\frac{J}{KL}, \frac{K}{JL}, \frac{L}{JK}\right) + 1\right)$, the maximum likelihood estimators for the covariance matrices $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Omega}_i$ in the TME model exist with probability 1.

Proposition 9 is a three-dimensional extension of Theorem 8 on page 14 of Roś et al. (2016).

Due to the similarity of the MLE for covariance matrices $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Omega}_i$ and $\boldsymbol{\Sigma}_\varepsilon$, $\boldsymbol{\Psi}_\varepsilon$, $\boldsymbol{\Omega}_\varepsilon$, the existence condition of the MLE for covariance matrices $\boldsymbol{\Sigma}_\varepsilon$, $\boldsymbol{\Psi}_\varepsilon$, $\boldsymbol{\Omega}_\varepsilon$ can be obtained accordingly. We need to point out that the existence of MLE does not mean that the estimation has good identifiability and convergence to the global optimal solution. We will investigate the identifiability in Section 4.3.4 and convergence in Section 4.4.3.

4.3.4 Identifiability

The identifiability of a statistical model is essential because it ensures correct inference on model parameters. For the TME model, the identifiability is extremely complexed because it involves three aspects: (i) whether the fixed effects core tensor is identifiable; (ii) the identifiability of the Kronecker covariance structure, because

$\boldsymbol{\Omega}_i \otimes \boldsymbol{\Psi}_i = c\boldsymbol{\Omega}_i \otimes \frac{1}{c}\boldsymbol{\Psi}_i$ for any $c > 0$; (iii) and the identifiability of covariance matrices of random effects and residual errors, because $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T} + \boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\Psi}_i = \boldsymbol{B}_i^{(2)}\boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T} + \boldsymbol{\Psi}_{\varepsilon}$, $\boldsymbol{\Omega}_i = \boldsymbol{B}_i^{(3)}\boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T} + \boldsymbol{\Omega}_{\varepsilon}$. We will investigate the identifiability for each these aspects respectively.

Firstly, the identifiability of the TME model follows the identifiability definition of a linear mixed effects model (Demidenko 2013). If the TME model is determined by a family of distributions $\{\boldsymbol{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, as shown in Equation (25), which is parameterized by the vector $\boldsymbol{\theta}$, and $\boldsymbol{\Theta}$ is the parameter space. The model is identifiable on $\boldsymbol{\Theta}$ if $\boldsymbol{P}_{\boldsymbol{\theta}_1} = \boldsymbol{P}_{\boldsymbol{\theta}_2}$ implies that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. Identifiability is a necessary property for the adequacy of the TME model.

In the linear mixed effects model, the design matrix for fixed effects has to be full-ranked to realize unique estimation of fixed effects parameters. If $\left[\!\left[\boldsymbol{\mathcal{F}}_1; \boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}\right]\!\right] = \left[\!\left[\boldsymbol{\mathcal{F}}_2; \boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}\right]\!\right]$ implies $\boldsymbol{\mathcal{F}}_1 = \boldsymbol{\mathcal{F}}_2$, it means the fixed effects core tensor is indentifiable. To ensure the identifiability of fixed effects core tensor in the TME model, it must satisfy that the design matrices $\boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}$ have full rank. According to the property of Kronecker product, $\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)} \otimes \boldsymbol{A}_i^{(1)}$ has full rank.

The identifiability of the Kronecker covariance structure can be ensured by introducing additional constraints. Because the covariance matrices are positive definite, one kind of constraint can be fixing particular summations of the diagonal elements of $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$ or $\boldsymbol{\Omega}_i$ to be equal to 1; another possible constraint is to assume that the determinants

of two of the three covariance matrices are equal to 1. The constraints do not restrict the application of the TME model since the relative magnitude of the entries in the covariance matrix will be relevant to the key information we care about.

Based on $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T} + \boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\Psi}_i = \boldsymbol{B}_i^{(2)}\boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T} + \boldsymbol{\Psi}_{\varepsilon}$ and $\boldsymbol{\Omega}_i = \boldsymbol{B}_i^{(3)}\boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T} + \boldsymbol{\Omega}_{\varepsilon}$, we know that the covariance matrices of random effects and residual errors are not unique. In order to ensure identifiability in a similar way to the classical mixed effects model (Demidenko 2013), we need to ensure that the design matrices $\boldsymbol{B}_i^{(1)}, \boldsymbol{B}_i^{(2)}, \boldsymbol{B}_i^{(3)}$ have full rank and to specify the structure of the covariance matrices $\boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\Psi}_{\varepsilon}$, $\boldsymbol{\Omega}_{\varepsilon}$. In general, there are two ways to specify the structure of the covariance matrices: One way is to assume the covariance matrices $\boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\Psi}_{\varepsilon}$, $\boldsymbol{\Omega}_{\varepsilon}$ are diagonal matrices (corresponding to the independent noise). Notably, the covariance matrices for random effects $\boldsymbol{\Sigma}_{\mathrm{r}}$, $\boldsymbol{\Psi}_{\mathrm{r}}$, $\boldsymbol{\Omega}_{\mathrm{r}}$ and the total covariance matrices $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Omega}_i$ are not diagonal. Another way is to determine the noise pattern based on a phase-I analysis. For example the noise is found to have a signal dependent property and the noise parameters are consistent for different data acquisition time in Raman inspection of nanomanufacturing (Yue et al. 2017).

4.4 Double Flip-Flop Algorithm for Parameter Estimation of TME Model

4.4.1 Double Flip-Flop Algorithm

For existing maximum likelihood estimation of covariance matrices with Kronecker structure, a Flip-Flop algorithm has been proposed to update the estimation of several components sequentially and iteratively (Dutilleul 1999; Lu and Zimmerman 2004;

Manceur and Dutilleul 2013; Sakata 2016). We have derived the maximum likelihood estimators for the TME model in Section 4.3. In this section, we will propose a double Flip-Flop algorithm to implement the parameter estimation iteratively. The algorithm is shown in Table 9.

Table 9. Double Flip-Flop algorithm for the TME model

Double Flip-Flop Algorithm for the TME Model
Step 1: Initialize the core tensor $\widehat{\boldsymbol{\mathcal{F}}}^{\{0\}}$ *and design matrices* $\mathbf{A}_{\mathrm{i}}^{(1)}, \mathbf{A}_{\mathrm{i}}^{(2)}, \mathbf{A}_{\mathrm{i}}^{(3)}, \mathbf{B}_{\mathrm{i}}^{(1)}, \mathbf{B}_{\mathrm{i}}^{(2)}, \mathbf{B}_{\mathrm{i}}^{(3)}$, *and covariance matrices* $\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}^{\{0\}}, \widehat{\boldsymbol{\Psi}}_{\mathrm{r}}^{\{0\}}, \widehat{\boldsymbol{\Omega}}_{\mathrm{r}}^{\{0\}}, \widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{\{0\}}, \widehat{\boldsymbol{\Psi}}_{\varepsilon}^{\{0\}}, \widehat{\boldsymbol{\Omega}}_{\varepsilon}^{\{0\}}$. *Set the iteration number* $\mathrm{k} = 0$. ➢ *Calculate the mean response tensor* $\bar{\boldsymbol{\mathcal{Y}}}$, *and then use high-order orthogonal iteration (HOOI) to compute a rank-*(P_1, Q_1, R_1) *Tucker decomposition,* $\bar{\boldsymbol{\mathcal{Y}}} = \left[\!\left[\boldsymbol{\mathcal{F}}^{\{0\}}; \boldsymbol{A}_i^{(1)\{0\}}, \boldsymbol{A}_i^{(2)\{0\}}, \boldsymbol{A}_i^{(3)\{0\}} \right]\!\right]$. *The decomposed core tensor and factor matrices work as the initialized fixed effect core tensor and design matrices for fixed effects.* ➢ *Choose the design matrices for random effects* $\boldsymbol{B}_i^{(1)\{0\}}, \boldsymbol{B}_i^{(2)\{0\}}, \boldsymbol{B}_i^{(3)\{0\}}$ *as a subset of appropriate columns of the design matrices* $\boldsymbol{A}_i^{(1)\{0\}}, \boldsymbol{A}_i^{(2)\{0\}}, \boldsymbol{A}_i^{(3)\{0\}}$. ➢ *Compute the* $\widehat{\boldsymbol{\Sigma}}_i^{\{0\}}, \widehat{\boldsymbol{\Psi}}_i^{\{0\}}, \widehat{\boldsymbol{\Omega}}_i^{\{0\}}$.
Step 2: Increase iteration number k by 1.
Step 3: Keep $\widehat{\boldsymbol{\mathcal{F}}}^{\{k-1\}}$ *fixed and compute* $\widehat{\boldsymbol{\Sigma}}_i^{\{k\}}, \widehat{\boldsymbol{\Psi}}_i^{\{k\}}, \widehat{\boldsymbol{\Omega}}_i^{\{k\}}$. ➢ *Compute* $\widehat{\boldsymbol{\Sigma}}_i^{\{k\}}$ *by using Equation (31) (using* $\widehat{\boldsymbol{\Psi}}_i^{\{k-1\}}, \widehat{\boldsymbol{\Omega}}_i^{\{k-1\}}$*).* ➢ *Compute* $\widehat{\boldsymbol{\Psi}}_i^{\{k\}}$ *by using Equation (32) (using* $\widehat{\boldsymbol{\Sigma}}_i^{\{k\}}, \widehat{\boldsymbol{\Omega}}_i^{\{k-1\}}$*).* ➢ *Compute* $\widehat{\boldsymbol{\Omega}}_i^{\{k\}}$ *by using Equation (33) (using* $\widehat{\boldsymbol{\Sigma}}_i^{\{k\}}, \widehat{\boldsymbol{\Psi}}_i^{\{k\}}$*).*
Step 4: Keep $\widehat{\boldsymbol{\Sigma}}_i^{\{k\}}, \widehat{\boldsymbol{\Psi}}_i^{\{k\}}, \widehat{\boldsymbol{\Omega}}_i^{\{k\}}$ *fixed and compute* $\widehat{\boldsymbol{\mathcal{F}}}^{\{k\}}$ *by using Equation (30).*

Step 5: Iterate between steps 2 and 4 until convergence or until reaching a predetermined number of iterations K.
Step 6: Set the iteration number $t = 0$. Estimate $\widehat{\boldsymbol{R}}^{\{0\}}$ by the expectation mean in Equation (37).
Step 7: Increase iteration number t by 1.
Step 8: Keep $\widehat{\boldsymbol{\mathcal{F}}}^{\{k\}}$, $\widehat{\boldsymbol{\mathcal{R}}}^{\{t-1\}}$ fixed and compute $\widehat{\boldsymbol{\Sigma}}_{\varepsilon}$, $\widehat{\boldsymbol{\Psi}}_{\varepsilon}$, $\widehat{\boldsymbol{\Omega}}_{\varepsilon}$ considering given constraints. ➢ *Compute $\widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{\{t\}}$ by using Equation (38) (using $\widehat{\boldsymbol{\Psi}}_{\varepsilon}^{\{t-1\}}, \widehat{\boldsymbol{\Omega}}_{\varepsilon}^{\{t-1\}}$) and adjust it according to given constraints.* ➢ *Compute $\widehat{\boldsymbol{\Psi}}_{\varepsilon}^{\{t\}}$ by using Equation (39) (using $\widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{\{t\}}, \widehat{\boldsymbol{\Omega}}_{\varepsilon}^{\{t-1\}}$) and adjust it according to given constraints.* ➢ *Compute $\widehat{\boldsymbol{\Omega}}_{\varepsilon}^{\{t\}}$ by using Equation (40) (using $\widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{\{t\}}, \widehat{\boldsymbol{\Psi}}_{\varepsilon}^{\{t\}}$) and adjust it according to given constraints.*
Step 9: Keep $\widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{\{t\}}, \widehat{\boldsymbol{\Psi}}_{\varepsilon}^{\{t\}}, \widehat{\boldsymbol{\Omega}}_{\varepsilon}^{\{t\}}$ fixed and compute $\widehat{\boldsymbol{\mathcal{R}}}^{\{t\}}$ by using Equation (37).
Step 10: Iterate between steps 7 and 9 until convergence or until reaching a predetermined number of iterations T.

Note that the algorithm involves two iterative loops. The first one is related to the computation of the fixed effects and total covariance matrices, and the second one is relevant to the computation of covariance matrices of residual errors and random effects. Each loop follows the characteristic of a Flip-Flop algorithm, and that is why it is named after the double Flip-Flop algorithm.

4.4.2 Initialization of the Algorithm

Obtaining good initial values is important for parameter estimation in the TME model. For the initialization, we use high-order orthogonal iteration (HOOI) to compute a rank-(P_1, Q_1, R_1) Tucker decomposition. In the HOOI algorithm, a higher-order Singular

Value Decomposition (HOSVD) is applied to initialize the factor matrices, and a set of orthogonal constraints to ensure the core tensor is all-orthogonal. This improves the uniqueness of Tucker decomposition (Kolda and Bader 2009). It is typically a challenging task to determine the parameters P_1, Q_1, R_1. Currently, we try multiple runs to check the sparsity of core tensor and determine the suitable parameters P_1, Q_1, R_1. The design matrices $\boldsymbol{A}_i^{(1)\{0\}}, \boldsymbol{A}_i^{(2)\{0\}}, \boldsymbol{A}_i^{(3)\{0\}}$ are determined by the factor matrices from the Tucker decomposition. The design matrices for random effects $\boldsymbol{B}_i^{(1)\{0\}}, \boldsymbol{B}_i^{(2)\{0\}}, \boldsymbol{B}_i^{(3)\{0\}}$ can be chosen as a subset of appropriate columns of the design matrices $\boldsymbol{A}_i^{(1)\{0\}}, \boldsymbol{A}_i^{(2)\{0\}}, \boldsymbol{A}_i^{(3)\{0\}}$. While the columns are determined by possible random effects relevant to features of interest in a phase-I data analysis. In phase-I data analysis, a set of process data is gathered and analyzed all at once in a retrospective analysis, and the features of interest will be chosen by multiple trials.

4.4.3 Convergence of the Algorithm

Lu and Zimmerman (2004) have explored the convergence of a Flip-Flop algorithm. According to the paper (Lu and Zimmerman 2004), the likelihood function of successive iterations of a Flip-Flop algorithm cannot decrease. Provided $N \geq JKL$, the algorithm is guaranteed to converge. However, whether it converges to a MLE is not ensured because the space of the covariance matrices is not convex. An empirical study of the convergence is investigated in Section 4.5.1.

The most commonly used stopping criteria are ones that based on the relative change in either the covariance parameters between successive iterations or differences between successive log-likelihood functions. Considering all the covariance matrices, the

stopping criteria for the first loop are that the L_1 norms $\left\|\widehat{\boldsymbol{\Sigma}}_i^{\{k\}} - \widehat{\boldsymbol{\Sigma}}_i^{\{k-1\}}\right\|_1$, $\left\|\widehat{\boldsymbol{\Psi}}_i^{\{k\}} - \widehat{\boldsymbol{\Psi}}_i^{\{k-1\}}\right\|_1$, $\left\|\widehat{\boldsymbol{\Omega}}_i^{\{k\}} - \widehat{\boldsymbol{\Omega}}_i^{\{k-1\}}\right\|_1$ are simultaneously smaller than the thresholds. Similar stopping criteria are applied for the second loop, which means that $\left\|\widehat{\boldsymbol{\Sigma}}_\varepsilon^{\{t\}} - \widehat{\boldsymbol{\Sigma}}_\varepsilon^{\{t-1\}}\right\|_1$, $\left\|\widehat{\boldsymbol{\Psi}}_\varepsilon^{\{t\}} - \widehat{\boldsymbol{\Psi}}_\varepsilon^{\{t-1\}}\right\|_1$, $\left\|\widehat{\boldsymbol{\Omega}}_\varepsilon^{\{t\}} - \widehat{\boldsymbol{\Omega}}_\varepsilon^{\{t-1\}}\right\|_1$ are simultaneously smaller than the thresholds. For the asymptotic properties of the Flip-Flop type algorithm, please refer to the paper (Werner et al. 2008). We also investigate the asymptotic properties in Sections 4.5 through simulation and surrogated data analysis.

4.4.4 Computational Complexity of the Algorithm

Since the double Flip-Flop algorithm uses the HOOI algorithm to do initialization and then conducts two iterative Flip-Flop loops, we need to analyze the computational cost of this algorithm. For simplicity, we assume the dimensions $J = K = L, P_1 = Q_1 = R_1$, $P_2 = Q_2 = R_2$. In the initialization part, each iteration in HOOI involves six tensor-by-matrix products and three maximization problems, where the computational complexity for each HOOI iteration is $O(J^3P_1 + JP_1^4 + P_1^6)$ (Elden and Savas 2009). The computational complexity of step 3 is $O(NJ^5)$, while step 4 is $O(NP_1^9 + NJ^3P_1^3)$. Therefore the cost for each iteration of the first Flip-Flop loop is $O(NJ^5 + NP_1^9 + NJ^3P_1^3)$. Similarly, the computational complexities for step 8 and step 9 are $O(NJ^5)$ and $O(NJ^3P_2)$, respectively. Thus, the computation cost for each iteration of the second Flip-Flop loop is $O(NJ^5)$, which is dominated by step 8. The computational time will also be impacted by the iteration number. According to the simulation study in Section 4.5.1, the algorithm will converge quickly.

To show the computational advantage of the proposed TME model, we consider the conventional linear mixed effects model for vectorized responses (marked as vLME) (Galecki and Burzykowski 2013). After vectorization of the tensor responses, the dimension of each response becomes J^3. Thus the computational complexity of the vLME model is $O(NJ^9)$ (Lippert et al. 2011). It is much larger than the complexity of the TME model, which is $O(NJ^5 + NP_1^9 + NJ^3P_1^3)$.

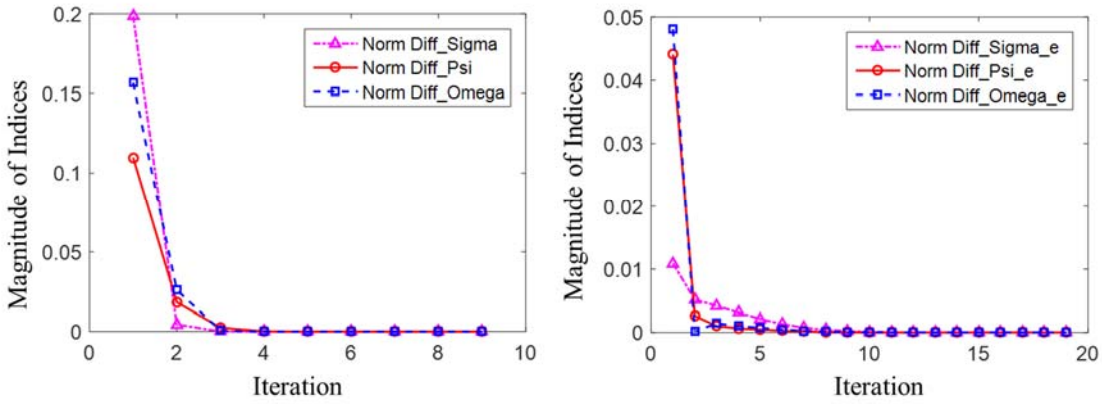
4.5 Numerical Analysis

4.5.1 Simulation Study

In this section, the performance of the iterative algorithm is evaluated through simulation studies. In order to simulate the response tensor with mixed effects, we generate the fixed effects tensor with dimension 30×5×5. The dimensions of core tensors for fixed effects and random effects are 8×3×3 and 3×2×2, respectively. The covariance matrices of random effects are generated from random symmetric positive definite matrices. Two covariance matrices of residual errors are generated by isotropic matrices (an isotropic matrix is an identity matrix multiplied by a positive number) with dimension 5×5 and another covariance matrix with dimension 30×30. 1000 response tensors are generated to test the performance. A computer with Intel Core i7-4500U processor and 8.00GB RAM is used to conduct the numerical analysis.

After we generate the dataset, we run the double Flip-Flop algorithm for parameter estimation of a tensor mixed effect model. For the performance of convergence, we test several convergence indices that are the divided L_1 norm of the difference between

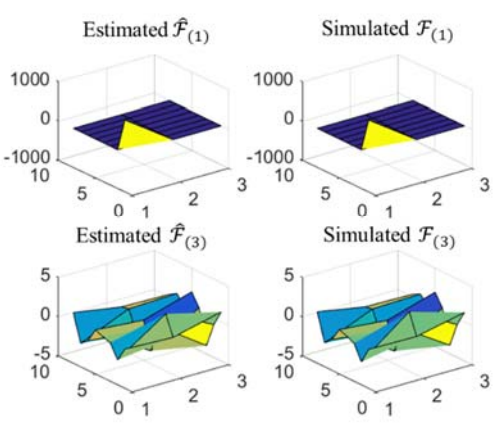
covariance matrices in two successive iterations, including $\left\|\widehat{\boldsymbol{\Sigma}}_i^{\{k\}} - \widehat{\boldsymbol{\Sigma}}_i^{\{k-1\}}\right\|_1 / J \cdot J$, $\left\|\widehat{\boldsymbol{\Psi}}_i^{\{k\}} - \widehat{\boldsymbol{\Psi}}_i^{\{k-1\}}\right\|_1 / K \cdot K$, $\left\|\widehat{\boldsymbol{\Omega}}_i^{\{k\}} - \widehat{\boldsymbol{\Omega}}_i^{\{k-1\}}\right\|_1 / L \cdot L$ for the first loop and $\left\|\widehat{\boldsymbol{\Sigma}}_\varepsilon^{\{t\}} - \widehat{\boldsymbol{\Sigma}}_\varepsilon^{\{t-1\}}\right\|_1 / J \cdot J$, $\left\|\widehat{\boldsymbol{\Psi}}_\varepsilon^{\{t\}} - \widehat{\boldsymbol{\Psi}}_\varepsilon^{\{t-1\}}\right\|_1 / K \cdot K$, $\left\|\widehat{\boldsymbol{\Omega}}_\varepsilon^{\{t\}} - \widehat{\boldsymbol{\Omega}}_\varepsilon^{\{t-1\}}\right\|_1 / L \cdot L$ for the second loop. We can see the convergence indices versus iterative histories in Figure 21. We can find that the convergence history is monotonic and fast.



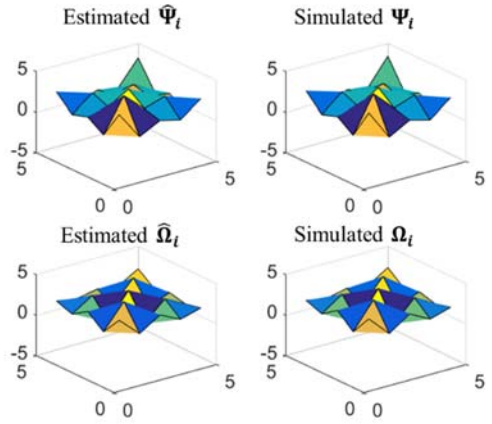
(a) Convergence of the first loop (b) Convergence of the second loop

Figure 21. Convergence of the iterative algorithm

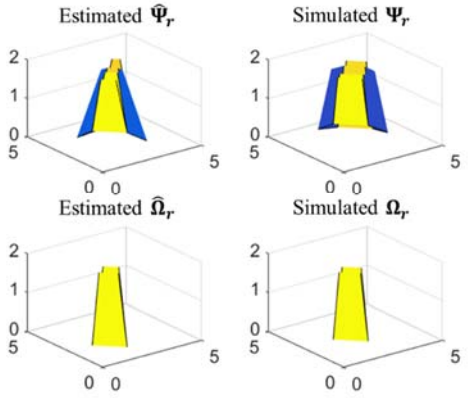
When we do the parameter estimation for the TME model, we will get the design matrices $\boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}$ first by a Tucker decomposition for the mean response tensor. The design matrices $\boldsymbol{B}_i^{(1)}, \boldsymbol{B}_i^{(2)}, \boldsymbol{B}_i^{(3)}$ are a subset of the design matrices $\boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}$. After convergence, we compare the estimated parameters with a sample size of 600 and the ones in the simulation model (underlying true parameters), including core tensor of fixed effects, total covariance matrices from the first loop and covariance matrices of residual errors from the second loop. The results are shown in Figure 22. We can see the estimations are quite consistent with the simulated parameters (underlying true parameters).



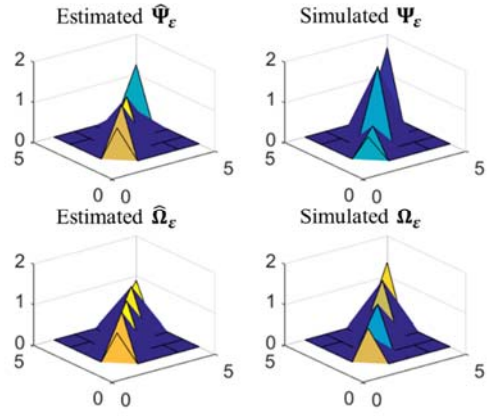
(a) Core tensor of fixed effects



(b) Covariance matrices $\widehat{\boldsymbol{\Psi}}_i$ and $\widehat{\boldsymbol{\Omega}}_i$



(c) Covariance matrices $\widehat{\boldsymbol{\Psi}}_r$ and $\widehat{\boldsymbol{\Omega}}_r$



(d) Covariance matrices $\widehat{\boldsymbol{\Psi}}_\varepsilon$ and $\widehat{\boldsymbol{\Omega}}_\varepsilon$

Figure 22. Comparison between estimated parameters and simulated parameters

In order to quantitatively evaluate the estimation accuracy, we introduce indices, including $\boldsymbol{D}_{\mathcal{F}}$, $\boldsymbol{D}_{\boldsymbol{\Sigma}_i}$, $\boldsymbol{D}_{\boldsymbol{\Psi}_i}$, $\boldsymbol{D}_{\boldsymbol{\Omega}_i}$ for the first loop, and $\boldsymbol{D}_{\boldsymbol{\Sigma}_\varepsilon}$, $\boldsymbol{D}_{\boldsymbol{\Psi}_\varepsilon}$, $\boldsymbol{D}_{\boldsymbol{\Omega}_\varepsilon}$ for the second loop. Where $\boldsymbol{D}_{\boldsymbol{X}} = \left\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\right\|_F / \|\boldsymbol{X}\|_F$, and $\widehat{\boldsymbol{X}} - \boldsymbol{X}$ denote the difference between estimated and true matrix/tensor. Moreover, $\boldsymbol{X} = \{\boldsymbol{\mathcal{F}}, \boldsymbol{\Sigma}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\Sigma}_\varepsilon, \boldsymbol{\Psi}_\varepsilon, \boldsymbol{\Omega}_\varepsilon\}$, and $\|\cdot\|_F$ denotes Frobenius norm. Furthermore, we introduce three indices for showing the convergence speed of different sample sizes. These indices are the iteration number, the time per iteration in the first loop, and the time per iteration in the second loop.

In order to explore the quantitative estimation accuracy and the asymptotic properties, we conduct the parameter estimation of the TME model for different sample sizes from 50 to 800. One hundred simulation runs are tested, where the mean and the standard deviation of the quantitative indices $\boldsymbol{D}_{\boldsymbol{X}}, \boldsymbol{X} = \{\boldsymbol{\mathcal{F}}, \boldsymbol{\Sigma}_i, \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\Sigma}_\varepsilon, \boldsymbol{\Psi}_\varepsilon, \boldsymbol{\Omega}_\varepsilon\}$ are calculated. The results are listed in Table 10. For the convergence speed, we notice that as the increase of sample size from 50 to 800, the average iteration number becomes smaller (from 9.32 to 6.05). While the average time per iteration increases from 0.62 seconds to 9.89 seconds in the first loop and from 0.22 seconds to 3.48 seconds in the second loop. For the quantitative estimation accuracy, the indices $\boldsymbol{D}_{\boldsymbol{\mathcal{F}}}, \boldsymbol{D}_{\boldsymbol{\Sigma}_i}, \boldsymbol{D}_{\boldsymbol{\Psi}_i}, \boldsymbol{D}_{\boldsymbol{\Omega}_i}, \boldsymbol{D}_{\boldsymbol{\Sigma}_\varepsilon}, \boldsymbol{D}_{\boldsymbol{\Psi}_\varepsilon}, \boldsymbol{D}_{\boldsymbol{\Omega}_\varepsilon}$ become smaller as the sample size increases. Which means that the estimated parameters are more accurate for the larger sample size. Of course, it costs more to obtain larger size of samples.

We compare our proposed TME method with three benchmark methods. The first is the tensor normal model with a structured mean (Nzabanita et al. 2015), which corresponds to the model that only considers fixed effects. We name it as the Tensor Fixed Effects (TFE) model and it is shown in Equation (41).

$$\boldsymbol{\mathcal{Y}}_i = \boldsymbol{\mathcal{F}} \times_1 \boldsymbol{A}_i^{(1)} \times_2 \boldsymbol{A}_i^{(2)} \times_3 \boldsymbol{A}_i^{(3)} + \boldsymbol{\mathcal{E}}_i \tag{41}$$

where the distribution of the residual errors tensor $\boldsymbol{\mathcal{E}}_i$ is $\boldsymbol{N}_{J,K,L}(\boldsymbol{\mathcal{O}}; \boldsymbol{\Sigma}_\varepsilon, \boldsymbol{\Psi}_\varepsilon, \boldsymbol{\Omega}_\varepsilon)$, and the noise covariance matrices along different dimensions are $\boldsymbol{\Sigma}_\varepsilon \in \mathbb{R}^{J\times J}, \boldsymbol{\Psi}_\varepsilon \in \mathbb{R}^{K\times K}, \boldsymbol{\Omega}_\varepsilon \in \mathbb{R}^{L\times L}$. The second benchmark method is the Tucker decomposition (TD) for the average tensor response. In this method, we do not consider that the residual errors follow the tensor normal distribution. The third benchmark method is the vectorized Linear Mixed-effects

(vLME) model, which conduct the linear mixed-effects model (Galecki and Burzykowski 2013) after vectorization of the tensor responses.

Table 10. Quantitative results for convergence speed and accuracy for different sample size

Sample size	Iteration number	$\boldsymbol{D_{\mathcal{F}}}$	$\boldsymbol{D_{\Sigma_i}}$	$\boldsymbol{D_{\Psi_i}}$	$\boldsymbol{D_{\Omega_i}}$	Time 1/s
50	9.21 (0.48)	0.0093 (0.0013)	0.9463 (0.0016)	0.0999 (0.0131)	0.0353 (0.0151)	0.63 (0.06)
80	8.69 (0.46)	0.0071 (0.0009)	0.9140 (0.0022)	0.0988 (0.0128)	0.0325 (0.0161)	0.99 (0.11)
100	8.12 (0.48)	0.0063 (0.0008)	0.8924 (0.0027)	0.0989 (0.0129)	0.0318 (0.0158)	1.26 (0.23)
200	7.02 (0.14)	0.0045 (0.0006)	0.7842 (0.0049)	0.0985 (0.0121)	0.0297 (0.0157)	2.47 (0.27)
400	6.71 (0.45)	0.0033 (0.0005)	0.5682 (0.0094)	0.0981 (0.0122)	0.0280 (0.0158)	4.94 (0.50)
600	6.27 (0.45)	0.0027 (0.0004)	0.3525 (0.0143)	0.0982 (0.0124)	0.0277 (0.0162)	7.42 (0.77)
800	6.07 (0.26)	0.0023 (0.0003)	0.1380 (0.0181)	0.0982 (0.0126)	0.0277 (0.0162)	9.86 (1.05)

Sample size	$\boldsymbol{D_{\Sigma_\varepsilon}}$	$\boldsymbol{D_{\Psi_\varepsilon}}$	$\boldsymbol{D_{\Omega_\varepsilon}}$	Time 2/s
50	0.9596 (0.3692)	0.2900 (0.0298)	0.3586 (0.0344)	0.27 (0.03)
80	0.5566 (0.2811)	0.2685 (0.0234)	0.3327 (0.0269)	0.41 (0.05)
100	0.3966 (0.1835)	0.2577 (0.0215)	0.3188 (0.0240)	0.49 (0.06)
200	0.1827 (0.0355)	0.2313 (0.0183)	0.2876 (0.0191)	0.99 (0.13)
400	0.1249 (0.0205)	0.2189 (0.0166)	0.2736 (0.0180)	1.99 (0.26)
600	0.1100 (0.0182)	0.2157 (0.0165)	0.2696 (0.0178)	2.97 (0.34)
800	0.1033 (0.0172)	0.2135 (0.0164)	0.2678 (0.0175)	3.98 (0.46)

In the simulation, we used the tensor toolbox from the Sandia National Laboratories (Bader, et al. 2015) when writing codes for the TME, the TFE and the TD. We used the fitlmematrix function in Matlab to conduct the vLME model. In order to evaluate the performance of the proposed TME model and benchmark methods, we use the mean square error for each sample denoted as $\text{MSE}_i = \left\|\boldsymbol{\mathcal{Y}}_i - \widehat{\boldsymbol{\mathcal{Y}}}_i\right\|_F^2 / JKL$ with $i = 1,\cdots,N$. The mean and standard deviation for MSE are calculated for different sample sizes and presented in Table 11. The time in Table 11 denotes the total running time of the corresponding models.

The results of mean and standard deviation for MSE and computational time in different sample sizes are shown in Table 11. For the general pattern, as the sample size increases, the mean of MSE tends to become smaller for all those methods. This is because the quantitative estimation accuracy is low when the sample size is small. When the sample size is 50, 80, or 100, the MSE of the proposed TME model is larger than that of the TFE model and the TD model. The reason is that the sample size is lower than the number of unknown parameters needed to be estimated, which is 152 in this simulation example. Therefore the parameter estimation is not accurate. It indicates that if the sample size is low, the error from parameter estimation will significantly reduce the effectiveness of the model that considers random effects. Hence, it is better to only consider the fixed effects. When the sample size is larger than 200, the proposed TME model outperforms the TFE model and the TD model with respect to MSE. This is especially true when the sample size is comparable or larger than the total dimensions ($J \cdot K \cdot L$ in this example). The reason is that the TME model considers not only the fixed effects, but also the random effects. Additionally, the computational time of the TME model, the TFE model, and the TD model are comparable.

Table 11. Comparison of mean square error in the TME model and benchmark methods

Sample size	TME		Benchmark 1 (TFE)		Benchmark 2 (TD)		Benchmark 3 (vLME)	
	MSE	Time	MSE	Time	MSE	Time	MSE	Time
50	67.94 (38.29)	9.29	12.47 (1.27)	5.97	12.39 (1.27)	1.69	9.46 (0.71)	188.17
80	26.25 (9.41)	15.24	12.42 (1.28)	8.20	12.36 (1.27)	2.44	9.38 (0.76)	338.60
100	18.37 (4.20)	18.08	12.39 (1.30)	11.10	12.35 (1.29)	2.86	9.35 (0.75)	350.55
200	11.97 (1.29)	22.82	12.36 (1.40)	17.37	12.34 (1.40)	5.75	9.36 (0.77)	1006.86
400	10.52 (0.92)	36.95	12.36 (1.46)	29.57	12.35 (1.47)	11.80	9.32 (0.77)	2188.93
600	10.22 (0.85)	58.13	12.39 (1.43)	46.93	12.38 (1.43)	18.35	9.35 (0.78)	2602.00
800	10.06 (0.82)	85.55	12.32 (1.42)	70.63	12.32 (1.42)	26.06	9.34 (0.78)	3754.12

By comparing the TME model and the vLME model, we find that the vLME model has a lower MSE than the TME model. The reason is that after vectorization, the vLME model tries to capture the random effects from the whole variables, without considering the constraints of the data structure. For the data with tensor structure, we assert that the vLME model will result in overfitting because it ignores the tensor structure. Meanwhile, the TME model can capture the complex correlation structure in different dimensions. Furthermore, the vLME model has significantly larger computational cost than the TME model, as discussed in Section 4.4.4. As shown in Table 11, the computational time of the vLME model increases from 188.17 seconds to 3754.12 seconds when the sample size increases from 50 to 800. The vLME is not efficient enough for in-line process monitoring and quality assessment. Also, when we run the simulation, it results in the (out of memory) error when the sample size is larger than 1000. The reason is that the memory required by

the vLME model is very large and the memory cost scales with the square of the sample size.

4.5.2 Surrogated Data Analysis of Raman Mapping

In this section, the performance of the TME model is evaluated through the surrogated Raman mapping data from a real CNTs buckypaper fabrication process. The setup of in-line Raman spectroscopy is shown in Figure 7. In the experimental setup, Near Infra-Red (NIR) laser with a wavelength of 785nm and a laser output power of 150mW were used to eliminate the effect of ambient light. A low magnification lens was used to achieve a larger focus tolerance.

Raman mapping data have been collected from multiple rectangular zones, and the Raman data from each zone corresponds to one tensor. One Raman mapping tensor is shown in Figure 20. Red dots represent measurement points, and there is a Raman spectrum in each measurement point. The mean response tensor is computed, and Tucker decomposition is conducted to obtain the design matrices. The dimension of the response tensor is 256×5×5. The dimensions of core tensor of fixed effects and random effects are 8×3×3 and 4×2×2, respectively. The covariance matrices of random effects are generated by weighted summations of diagonal matrices with random values and identity matrices. Two covariance matrices of residual errors are generated by the identity matrices with dimension 5×5 and another covariance matrix with dimension 256×256 that is diagonal with a given signal-dependent noise from experimental data.

After generating the surrogated Raman mapping data, the proposed TME is applied to extract different components including fixed effects, random effects, and signal-

dependent noise. To explore the quantitative estimation accuracy, the same indices $\boldsymbol{D}_{\mathrm{f1}}, \boldsymbol{D}_{\boldsymbol{\Sigma}_i}, \boldsymbol{D}_{\boldsymbol{\Psi}_i}, \boldsymbol{D}_{\boldsymbol{\Omega}_i}, \boldsymbol{D}_{\boldsymbol{\Sigma}_\varepsilon}, \boldsymbol{D}_{\boldsymbol{\Psi}_\varepsilon}, \boldsymbol{D}_{\boldsymbol{\Omega}_\varepsilon}$ that were defined in Section 4.5 have been used to evaluate the results under different sample sizes. The results are shown in Table 12.

Table 12. Quantitative results for convergence speed and accuracy for surrogated Raman mapping data

Sample size	Iteration number	$\boldsymbol{D}_{\boldsymbol{\mathcal{F}}}$	$\boldsymbol{D}_{\boldsymbol{\Sigma}_i}$	$\boldsymbol{D}_{\boldsymbol{\Psi}_i}$	$\boldsymbol{D}_{\boldsymbol{\Omega}_i}$	Time 1/s	$\boldsymbol{D}_{\boldsymbol{\Sigma}_\varepsilon}$	$\boldsymbol{D}_{\boldsymbol{\Psi}_\varepsilon}$	$\boldsymbol{D}_{\boldsymbol{\Omega}_\varepsilon}$	Time 2/s
50	17	0.0039	0.9184	0.2031	0.2076	7.08	4.6351	0.3624	0.2261	3.55
80	13	0.0035	0.8693	0.2063	0.2158	11.94	4.3259	0.3455	0.2165	5.55
100	13	0.0031	0.8365	0.2079	0.2174	15.83	3.9043	0.3276	0.2087	6.91
200	9	0.0022	0.6735	0.2089	0.2190	30.84	0.2483	0.2945	0.1815	13.93
400	9	0.0015	0.3465	0.2103	0.2201	56.83	0.1449	0.2900	0.1792	28.62
600	7	0.0012	0.0702	0.2107	0.2206	90.73	0.1264	0.2908	0.1789	42.32

In comparison to Table 10, the results in Table 12 show similar asymptotic patterns, which means the estimated parameters become more accurate as the sample size increases. For example, when the sample size is 600, the indices $\boldsymbol{D}_{\boldsymbol{\mathcal{F}}}, \boldsymbol{D}_{\boldsymbol{\Sigma}_i}, \boldsymbol{D}_{\boldsymbol{\Psi}_i}, \boldsymbol{D}_{\boldsymbol{\Omega}_i}$ are as low as 0.0012, 0.0702, 0.2107 and 0.2206, respectively, which indicates the accuracy of the parameter estimation. However, the iteration number and time per iteration become larger for the same sample size due to the dimension increase from 30 to 256. Specifically, the computation time for 600 samples are 90.73 seconds per first loop and 42.32 per second loop.

4.5.3 Case Study

In this section, we show a real case study of the application of the proposed TME model. The setup of in-line Raman spectroscopy is shown in Figure 7. Similar to the surrogated data analysis, NIR laser with a wavelength of 785nm and a laser output power of 150mW were used to eliminate the effect of ambient light. The multi-walled CNTs buckypaper before alignment and after alignment are measured by Raman mapping technique. The scanning electron microscope (SEM) pictures for CNTs buckypaper are shown in Figure 23. Alignment was conducted by stretching with the different stretch ratio, including 0%, 20%, 35%, and 60%. When stretch ratio equals 0%, it is referring to the CNTs buckypaper before alignment. The matrices $\mathbf{\Psi}_r$ and $\mathbf{\Psi}_\varepsilon$ are associated with the correlation along the horizontal direction, while the matrices $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_\varepsilon$ are associated with the correlation along the vertical direction.

After running Raman mapping in a rectangular zone, 800 response tensors with dimension 256×3×4 are generated from each CNTs buckypaper sample. The proposed TME model is used to fit the datasets and the double flip-flop algorithm is conducted for parameter estimation. The covariance matrices $\mathbf{\Psi}_r$, $\mathbf{\Psi}_\varepsilon$, $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_\varepsilon$ for CNTs buckypaper with different alignment degree are summarized in Figure 24 and Figure 25. Figure 24(a) shows the range of diagonal entries in covariance matrices ($\mathbf{\Psi}_r$ and $\mathbf{\Psi}_\varepsilon$) along the horizontal direction; while Figure 24(b) shows the range of diagonal entries in covariance matrices ($\mathbf{\Omega}_r$ and $\mathbf{\Omega}_\varepsilon$) along the vertical direction. Figure 25 indicates changes of covariance coefficients as the stretch ratio increases. Figure 25(a) shows coefficients in $\mathbf{\Psi}_r$, and Figure 25(b) shows coefficients in $\mathbf{\Omega}_r$.

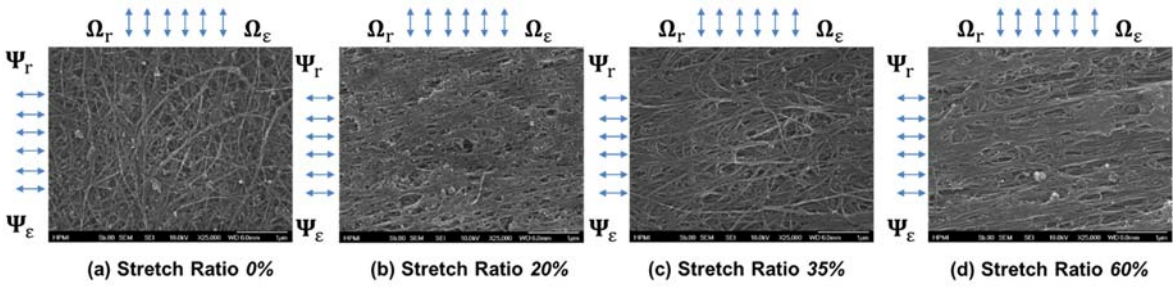


Figure 23. SEM pictures of CNTs buckypaper with different stretch ratios

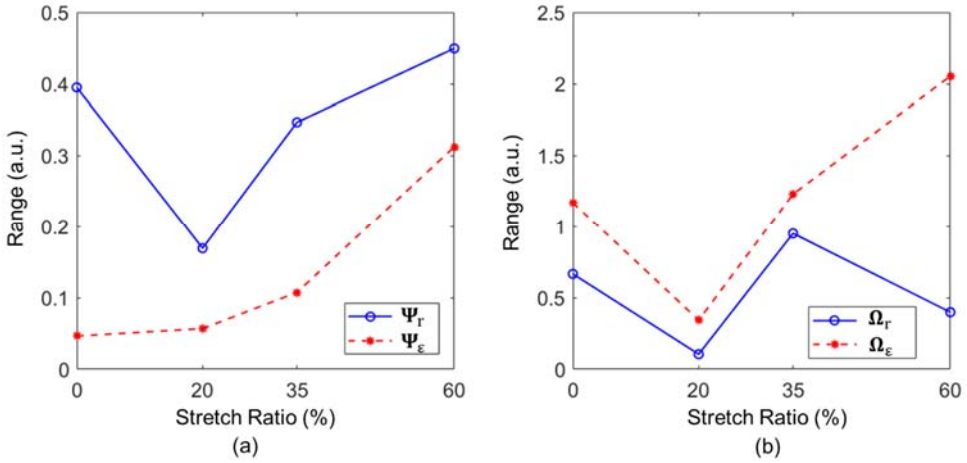


Figure 24. The range of diagonal entries in covariance matrices (a) Ψ_r and Ψ_ε, (b) Ω_r and Ω_ε for CNTs buckypaper with different alignment degree

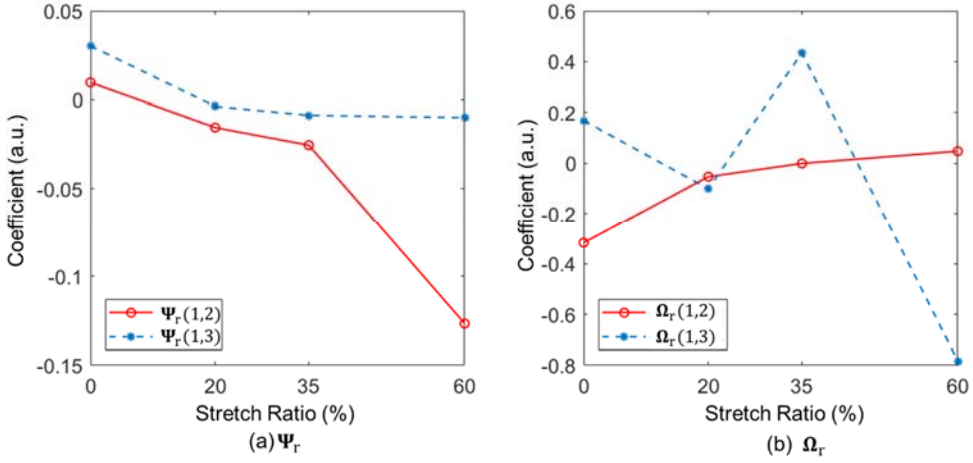


Figure 25. Covariance coefficients in covariance matrices (a) Ψ_r, (b) Ω_r for CNTs buckypaper with different alignment degree

We compare the covariance matrices along horizontal and vertical directions for each CNTs buckypaper sample. Considering the physical knowledge of CNTs buckypaper, we can provide the following remarks:

(i) For the random effects covariance along the horizontal direction, we observe that the coefficient $\boldsymbol{\Psi}_{\mathrm{r}}(1,2)$ tends to become negative after alignment and as the degree of alignment increases, the absolute magnitude becomes larger. The covariance coefficient $\boldsymbol{\Psi}_{\mathrm{r}}(1,2)$ changes from 0.0097 to -0.1267. It indicates that negative correlation along the alignment direction occurs, and the covariance coefficient changes with the alignment of CNTs buckypaper. This can be explained by the conservation of mass in a local zone. Alignment introduces systematic ridges and valleys in the microstructure pattern, and a ridge will be close to a valley, that indicates the negative correlation in the height. The height will impact the measurement distance between the laser head in Raman spectroscopy and the Raman mapping. The covariance coefficient in $\boldsymbol{\Psi}_{\mathrm{r}}(1,3)$ becomes negative after alignment, but the absolute magnitude becomes closer to zero. One physical interpretation is that after alignment, the distance between the first measurement line and the third measurement line becomes larger, and their correlation relationship becomes weaker.

(ii) For the random effects covariance along the vertical direction, coefficients $\boldsymbol{\Omega}_{\mathrm{r}}(1,2)$ becomes closer to zero, but the absolute magnitude of $\boldsymbol{\Omega}_{\mathrm{r}}(1,3)$ becomes larger. The physical interpretation is that as the stretch ratio increases, the high-frequency surface roughness becomes smaller, while the low-frequency surface roughness becomes larger.

(iii) For the range of diagonal entries in covariance matrices of Raman mapping, the change of $\boldsymbol{\Psi}_r$, $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_\varepsilon$ for CNTs buckypaper with different degrees of alignment are quite random. However, $\boldsymbol{\Psi}_\varepsilon$ has a larger quantitative difference between the maximum entry and the minimum one after alignment. Without alignment, the diagonal coefficients range is 0.0464, while after alignment with a stretch ratio of 60%, the range becomes as large as 0.3121. Which means that for different measurement lines along the alignment direction, the variability becomes larger. This makes sense because the alignment creates systematic ridges and valleys along the alignment directions. We can use this index to quantify the degree of alignment.

In summary, based on the covariance matrices from the TME model, we can quantify the influence of alignment based on the range of diagonal entries in $\boldsymbol{\Psi}_\varepsilon$ and covariance coefficients $\boldsymbol{\Psi}_r(1,2)$. The quantitative changes after alignment can be interpreted by engineering knowledge.

4.6 Conclusion

In this chapter, we proposed a novel TME model that effectively and efficiently explores the fixed effects and random effects inherent to the data in tensor domain. The advantages of this model include (i) its capability to handle multilevel hierarchical data; (ii) its ability to take complexed association structures, including correlation along different dimensions, into consideration; (iii) analyzing the mixed effects for the high-dimensional datasets. The proposed TME model can be viewed as a logical extension from a vector/matrix-valued mixed effects model to an array-valued mixed effects model. The proposed TME model is applied in the nanomanufacturing inspection. Moreover, the TME

model can be applied to provide potential solutions for a family of tensor data analytics with mixed effects, such as problems in the research fields of multimodality imaging analysis, chemometrics, neuroimaging, multichannel signal processing, etc.

For the TME model, the distribution of response tensors and its k-mode matricization were explored. We also derived the log-likelihood function for the TME model. Maximum likelihood estimators for fixed effect core tensor and covariance matrices were derived. Existence of the MLE and identifiability of the TME model were illustrated. Moreover, an iterative double Flip-Flop algorithm has been developed for parameter estimation, and the initialization and convergence criteria have been discussed. The computational complexity of the Flip-Flop algorithm has been derived. The TME model was shown to outperform vectorized LME model from a computational complexity perspective. By simulation and surrogated data analysis, we found that the algorithm can realize very quick convergence. The iteration number becomes smaller and time per iteration becomes longer as the sample size increases. In addition, the asymptotic property was investigated in the simulation and surrogate data analysis. The estimation accuracy of total covariance matrices and covariance matrices for the error terms improved as the sample size increased. In the simulation study, we also find that the TME model outperforms two benchmark methods which do not consider random effects (the TFE model and the TD model) when the sample size is larger than the dimensions of response tensor. The TME model outperforms the vLME model when it comes to computational efficiency. Furthermore, in the case study, the influence of alignment of CNTs buckypaper is quantified by the covariance matrices along different dimensions.

CHAPTER 5. SUMMARY AND FUTURE RESEARCH

5.1 Summary of Original Contributions

This dissertation contributes to the area of *System Informatics and Control* to develop systematic methodologies for data analytics and process monitoring in the continuous nanomanufacturing process. This is an interdisciplinary research area. We have developed novel methodologies by integrating advanced statistics, machine learning and engineering domain knowledge. The proposed methodologies mainly focus on data denoising for quality monitoring, data decomposition for profile detection, and high-dimensional data analytics for nanomanufacturing inspection. The original contributions of this dissertation include the following aspects:

- *A novel generalized wavelet shrinkage (GWS) method was proposed to remove the signal-dependent noise efficiently in the signals*. The main idea of the GWS method is to adaptively determine thresholds for shrinkage. Conventional wavelet shrinkage applies level-dependent thresholds to shrinkage wavelet coefficients since the variance cannot be estimated in every coefficient. The level-dependent threshold is suitable for a signal model with white noise. However, for signal-dependent noise in Raman spectra, the level-dependent threshold is not optimal. The GWS method is designed to adjust the threshold individually for each coefficient, which outperforms the level-dependent threshold. Many popular denoising methods assume the noise is white noise, which does not make sense for many signals in engineering systems. The proposed GWS method open a door for denoising in-line signals with signal-

dependent property, which lays a solid foundation for data-based process monitoring and quality improvement.

- *A new algorithm named after "penalized mixed-effects decomposition (PMD)" was developed to separate in-line profiles into four components: fixed effects, normal effects, defective effects, and noise*. The proposed PMD exploits a regularized high-dimensional regression with linear constraints. An optimization algorithm based on the accelerated proximal gradient (APG) is developed to do parameter estimation efficiently. The PMD algorithm can generate low-dimensional features from the in-line high-dimensional signals. The separated fixed effects coefficients, normal effects coefficients, and defective effects coefficients can be used to monitor quality characteristics such as fabrication consistency, uniformity, and defect information, respectively. The PMD algorithm can realize deeper data decomposition efficiently with ensuring identifiability, which makes it more attractive than the benchmark methods in the data decomposition and feature extraction domain.
- *A novel tensor mixed-effects (TME) model was developed to analyze massive high-dimensional data with complex correlation structure*. The proposed TME model can (i) separate fixed effects and random effects in a tensor domain; (ii) explore the correlations along different dimensions; and (iii) realize efficient parameter estimation by a proposed double Flip-Flop algorithm. The conventional mixed-effects model treats multivariate data as a vector or matrix, which is insufficient for high-dimensional data analysis. The current tensor decompositions such as CP decomposition or Tucker decomposition do not consider mixed effects and complex correlation structure in the data. Our main contribution is to extend the mixed-effects

model to multi-array (tensor) domain, and provide this new tool to explore the complex correlation structure for high-dimensional data analytics.

The proposed methodologies have laid a solid foundation for engineering-driven data analytics for nanomanufacturing. The GWS method allows for a significant reduction of Raman data acquisition time without much loss of S/N ratio, which dramatically improves in-line monitoring and control for continuous nanomanufacturing processes. The PMD algorithm enables monitoring of fabrication consistency, uniformity and defect information simultaneously. The TME model can be used to quantify the alignment degree of CNTs buckypaper.

Furthermore, the proposed methodologies can be applied to not only nanomanufacturing, but also other complex engineering systems. The GWS method can be used to denoising in-line signals with the signal-dependent property; the PMD method is suitable for other datasets with several characteristics (multichannel, mixture of physical information, signal-dependent noise); the TME model can be applied to other datasets with tensor structure, mixed effects and complex spatial and temporal correlations.

5.2 Future Research

There are several important topics to be explored for further development of process monitoring and quality improvement in the continuous nanomanufacturing. For example, the developed methodologies are mainly focused on in-line monitoring and detection of nanomanufacturing. In the multi-stage CNTs buckypaper fabrication process, there are many control factors and noise factors from different stages that may impact the final product quality. An automatic process control needs to be developed to improve the quality

of CNTs buckypaper. Moreover, supervisory control strategy need to be investigated to adaptively adjust the control parameters under various scenarios to realize the optimization of multistage nanomanufacturing process.

From the point-of-view of data analytics methodologies, my future research involves the development of engineering-driven data analytics methodologies for complex systems. The large-size, high-dimensional, heterogeneous data and extraordinary complex system structures make this a rich topic with numerous potential avenues of investigation. In particular, my research agenda is not limited to develop pure data-driven models or machine learning techniques, but also encompasses complex system structures and domain knowledge. For example, one future research direction is physical-based data decomposition, which is a further development of the PMD and TME methods. I would like to continue my research to develop a unified deterministic/stochastic data decomposition framework. Deterministic decomposition captures global geometric, topological, or algebraic characteristics of data, and it is a machine-learning type of decomposition. Stochastic decomposition establishes generative descriptions of data, which is associated with the advanced statistical models. Future research topics include theoretical exploration of bridging deterministic decomposition and stochastic decomposition, engineering-driven machine learning with explicit interpretations from the engineering knowledge.

APPENDIX A. PROOFS

This appendix illustrates the proofs of the propositions in Chapter 3 and Chapter 4.

A.1 Proof of the Proposition 1

Given $\boldsymbol{\delta}_{ij}$, $\boldsymbol{\theta}_{ij}$ can be derived by the optimization problem below.

$$\boldsymbol{\theta}_{ij} = \underset{\boldsymbol{\theta}_{ij}}{\operatorname{argmin}} \ (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}) + \gamma \left|\boldsymbol{\delta}_{ij}\right|_1$$

$$\text{subject to.} \quad \boldsymbol{B}_L \preccurlyeq \boldsymbol{W}\boldsymbol{\theta}_{ij} \preccurlyeq \boldsymbol{B}_U$$

where $\boldsymbol{z}_{ij} = \boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}_{\boldsymbol{a}}\boldsymbol{\delta}_{ij}$.

Since the $\boldsymbol{\Omega}$ is diagonal matrix with each entry is positive, we can solve the optimization component-wise by univariate quadratic programming with linear constraints. (Note: If $\boldsymbol{\Omega}$ is not diagonal, the optimization can be solved by the quadratic programming with linear constraints. However, the computation speed will be slower.) Considering the orthogonal wavelet basis matrix $\boldsymbol{W}$, the problem can be solved by a closed form solution,

$$\boldsymbol{\theta}_{ij}^{*} = \boldsymbol{W}^T \boldsymbol{z}_{ij} \circ \boldsymbol{I}\left(\boldsymbol{B}_L \preccurlyeq \boldsymbol{z}_{ij} \preccurlyeq \boldsymbol{B}_U\right) + \boldsymbol{W}^T \boldsymbol{B}_L \circ \boldsymbol{I}\left(\boldsymbol{z}_{ij} \preccurlyeq \boldsymbol{B}_L\right) + \boldsymbol{W}^T \boldsymbol{B}_U \circ \boldsymbol{I}\left(\boldsymbol{z}_{ij} \succcurlyeq \boldsymbol{B}_L\right)$$

It can be showed that

$$\boldsymbol{\delta}_{ij} = \underset{\boldsymbol{\delta}_{ij}}{\operatorname{argmin}} \ F\left(\boldsymbol{\delta}_{ij}\right) = (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{z}_{ij} - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*}) + \gamma \left|\boldsymbol{\delta}_{ij}\right|_1 \quad ,$$

which is the weighted LASSO formulation.

A.2 Proof of the Proposition 2

It is readily to show that $\boldsymbol{\Omega^{-1}}$ is a diagonal, positive definite matrix with all diagonal entries equal to the variance of Raman intensity corresponds to the related Raman shift. Thus, the spectrum norm of matrix $\boldsymbol{\Omega^{-1}}$ is the maximum eigenvalue of the matrix $\boldsymbol{\Omega^{-1}}$, which is $1/\min(\mathrm{diag}(\boldsymbol{\Omega}))$.

$$\|\boldsymbol{\Omega^{-1}}\| = 1/\min(\mathrm{diag}(\boldsymbol{\Omega})).$$

We can write down the gradient function

$$\begin{aligned}\nabla f(\boldsymbol{a}) &= \nabla(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W_a a} - \boldsymbol{W\theta}_{ij}^*)^T \boldsymbol{\Omega}^{-1}\left(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W_a a} - \boldsymbol{W\theta}_{ij}^*\right) \\ &= -2\boldsymbol{W_a^T}\,\boldsymbol{\Omega}^{-1}\left(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W_a a} - \boldsymbol{W\theta}_{ij}^*\right)\end{aligned}$$

Then we can get the difference between the gradient function for $\boldsymbol{a}$ and the one for $\boldsymbol{b}$ satisfies

$$\begin{aligned}\|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\| &= \|2\boldsymbol{W_a^T}\,\boldsymbol{\Omega}^{-1}\boldsymbol{W_a}(\boldsymbol{a} - \boldsymbol{b})\| \le \|2\boldsymbol{W_a^T}\,\boldsymbol{\Omega}^{-1}\boldsymbol{W_a}\|_2 \cdot \|\boldsymbol{a} - \boldsymbol{b}\| \\ &\le 2\|\boldsymbol{\Omega}^{-1}\|\,\|\boldsymbol{W_a}\|_2^2 \cdot \|\boldsymbol{a} - \boldsymbol{b}\| \\ &= \frac{2}{\min(\mathrm{diag}(\boldsymbol{\Omega}))}\,\|\boldsymbol{W_a}\|_2^2 \cdot \|\boldsymbol{a} - \boldsymbol{b}\|\end{aligned}$$

Thus $f(\cdot)$ is Lipchiz continuous, which implies that there is a constant L that makes gradient function $\nabla f(\cdot)$ satisfing $\|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\| \le L\|\boldsymbol{a} - \boldsymbol{b}\|$ for any $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{R}$ with $L = \frac{2}{\min(\mathrm{diag}(\boldsymbol{\Omega}))}\,\|\boldsymbol{W_a}\|_2^2$.

A.3 Proof of the Proposition 3

The proximal gradient algorithm for the penalized mixed-effects decomposition problem in Equation (22), given by $\boldsymbol{\delta_{ij}^{(k)}} = \underset{\boldsymbol{\delta}_{ij}}{\operatorname{argmin}} \left\{ f\left(\boldsymbol{\delta_{ij}^{(k-1)}}\right) + \langle \boldsymbol{\delta}_{ij} - \boldsymbol{\delta_{ij}^{(k-1)}}, \nabla f\left(\boldsymbol{\delta_{ij}^{(k-1)}}\right) \rangle + \frac{L}{2} \left\| \boldsymbol{\delta}_{ij} - \boldsymbol{\delta_{ij}^{(k-1)}} \right\|^2 + \gamma \left| \boldsymbol{\delta}_{ij} \right|_1 \right\}$ has a closed-form solution in each iteration k, in the form of a soft-thresholding function as follows:

$$\boldsymbol{\delta_{ij}^{(k)}} = S_{\frac{\gamma}{L}}\left(\boldsymbol{\delta_{ij}^{(k-1)}} + \frac{2}{L}\boldsymbol{W_a^T}\,\boldsymbol{\Omega}^{-1}\left(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*} - \boldsymbol{W_a}\boldsymbol{\delta_{ij}^{(k-1)}}\right)\right) \quad \text{with} \quad L = \frac{2}{\min(\operatorname{diag}(\boldsymbol{\Omega}))}\,\|\boldsymbol{W_a}\|_2^2,$$

where $S_{\frac{\gamma}{L}}(x) = \operatorname{sgn}(x)(|x| - \frac{\gamma}{L})_+$ is the soft-thresholding operator, and sgn(x) is the sign function and $\mathrm{x}_+ = \max(x, 0)$.

To prove the proposal above, we write down the gradient function:

$$\nabla f(\boldsymbol{a}) = -2\boldsymbol{W_a^T}\,\boldsymbol{\Omega}^{-1}\left(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*} - \boldsymbol{W_a}\boldsymbol{a}\right)$$

$$\nabla f\left(\boldsymbol{\delta_{ij}^{(k-1)}}\right) = -2\boldsymbol{W_a^T}\,\boldsymbol{\Omega}^{-1}\left(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*} - \boldsymbol{W_a}\boldsymbol{\delta_{ij}^{(k-1)}}\right)$$

For each iteration given by $(k-1)$ step,

$$\boldsymbol{\delta_{ij}^{(k)}} = \underset{\boldsymbol{\delta}_{ij}}{\operatorname{argmin}} \left\{ f\left(\boldsymbol{\delta_{ij}^{(k-1)}}\right) + \langle \boldsymbol{\delta}_{ij} - \boldsymbol{\delta_{ij}^{(k-1)}}, \nabla f\left(\boldsymbol{\delta_{ij}^{(k-1)}}\right) \rangle + \frac{L}{2} \left\| \boldsymbol{\delta}_{ij} - \boldsymbol{\delta_{ij}^{(k-1)}} \right\|^2 + \gamma \left| \boldsymbol{\delta}_{ij} \right|_1 \right\}$$

$$= \underset{\boldsymbol{\delta}_{ij}}{\operatorname{argmin}} \left\{ \frac{1}{2L} \left\| \nabla f\left(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}\right) \right\| + \langle \boldsymbol{\delta}_{ij} - \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}, \nabla f\left(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}\right) \rangle + \frac{L}{2} \left\| \boldsymbol{\delta}_{ij} - \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} \right\|^2 + \gamma \left| \boldsymbol{\delta}_{ij} \right|_1 \right\}$$

$$= \underset{\boldsymbol{\delta}_{ij}}{\operatorname{argmin}} \left\{ \frac{L}{2} \left\| \boldsymbol{\delta}_{ij} - \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} + \frac{1}{L} \nabla f\left(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)}\right) \right\|^2 + \gamma \left| \boldsymbol{\delta}_{ij} \right|_1 \right\}$$

$$= \underset{\boldsymbol{\delta}_{ij}}{\operatorname{argmin}} \left\{ \frac{L}{2} \left\| \boldsymbol{\delta}_{ij} - \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} - \frac{2}{L} \boldsymbol{W}_{\boldsymbol{a}}^{\boldsymbol{T}} \boldsymbol{\Omega}^{-1} \left(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*} - \boldsymbol{W}_{\boldsymbol{a}} \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} \right) \right\| + \gamma \left| \boldsymbol{\delta}_{ij} \right|_1 \right\}$$

$= S_{\frac{\gamma}{L}}(\boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} + \frac{2}{L} \boldsymbol{W}_{\boldsymbol{a}}^{\boldsymbol{T}} \boldsymbol{\Omega}^{-1} \left(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i - \boldsymbol{W}\boldsymbol{\theta}_{ij}^{*} - \boldsymbol{W}_{\boldsymbol{a}} \boldsymbol{\delta}_{\boldsymbol{ij}}^{(k-1)} \right))$ which is the exact soft thresholding solution.

A.4 Proof of the Proposition 4

Proof:

For simplification, we define $\widetilde{\boldsymbol{\mathcal{F}}} = \left[\!\left[\boldsymbol{\mathcal{F}}; \boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)} \right]\!\right]$, $\widetilde{\boldsymbol{\mathcal{R}}}_i = \left[\!\left[\boldsymbol{\mathcal{R}}_i; \boldsymbol{B}_i^{(1)}, \boldsymbol{B}_i^{(2)}, \boldsymbol{B}_i^{(3)} \right]\!\right]$

$$\boldsymbol{\mathcal{R}}_i \backsim \boldsymbol{N}_{P_2,Q_2,R_2}(\boldsymbol{\mathcal{O}}; \boldsymbol{\Sigma}_{\mathrm{r}}, \boldsymbol{\Psi}_{\mathrm{r}}, \boldsymbol{\Omega}_{\mathrm{r}}) \text{ if } \mathrm{vec}(\boldsymbol{\mathcal{R}}_i) \backsim \boldsymbol{N}_{P_2Q_2R_2}(\mathrm{vec}(\boldsymbol{\mathcal{O}}), \boldsymbol{\Omega}_{\mathrm{r}} \otimes \boldsymbol{\Psi}_{\mathrm{r}} \otimes \boldsymbol{\Sigma}_{\mathrm{r}})$$

$$\boldsymbol{\mathcal{Y}}_i = \boldsymbol{\mathcal{F}} \times_1 \boldsymbol{A}_i^{(1)} \times_2 \boldsymbol{A}_i^{(2)} \times_3 \boldsymbol{A}_i^{(3)} + \boldsymbol{\mathcal{R}}_i \times_1 \boldsymbol{B}_i^{(1)} \times_2 \boldsymbol{B}_i^{(2)} \times_3 \boldsymbol{B}_i^{(3)} + \boldsymbol{\mathcal{E}}_i$$

We check the distribution of random effects,

$$\mathrm{vec}\left(\boldsymbol{\mathcal{R}}_i \times_1 \boldsymbol{B}_i^{(1)} \times_2 \boldsymbol{B}_i^{(2)} \times_3 \boldsymbol{B}_i^{(3)} \right) = \left(\boldsymbol{B}_i^{(3)} \otimes \boldsymbol{B}_i^{(2)} \otimes \boldsymbol{B}_i^{(1)} \right) \mathrm{vec}(\boldsymbol{\mathcal{R}}_i)$$

$$\backsim \boldsymbol{N}_{IJK}\left(\left(\boldsymbol{B}_i^{(3)} \otimes \boldsymbol{B}_i^{(2)} \otimes \boldsymbol{B}_i^{(1)} \right) \mathrm{vec}(\boldsymbol{\mathcal{O}}), \left(\boldsymbol{B}_i^{(3)} \otimes \boldsymbol{B}_i^{(2)} \otimes \boldsymbol{B}_i^{(1)} \right) (\boldsymbol{\Omega}_{\mathrm{r}} \otimes \boldsymbol{\Psi}_{\mathrm{r}} \otimes \boldsymbol{\Sigma}_{\mathrm{r}}) \left(\boldsymbol{B}_i^{(3)} \otimes \boldsymbol{B}_i^{(2)} \otimes \boldsymbol{B}_i^{(1)} \right)^T \right)$$

$$\backsim \boldsymbol{N}_{IJK}\left(\mathrm{vec}(\boldsymbol{\mathcal{O}}), \left(\boldsymbol{B}_i^{(3)} \boldsymbol{\Omega}_{\mathrm{r}} \boldsymbol{B}_i^{(3)^T} \right) \otimes \left(\boldsymbol{B}_i^{(2)} \boldsymbol{\Psi}_{\mathrm{r}} \boldsymbol{B}_i^{(2)^T} \right) \otimes \left(\boldsymbol{B}_i^{(1)} \boldsymbol{\Sigma}_{\mathrm{r}} \boldsymbol{B}_i^{(1)^T} \right) \right).$$

Then $\widetilde{\boldsymbol{\mathcal{R}}}_i = \boldsymbol{\mathcal{R}}_i \times_1 \boldsymbol{B}_i^{(1)} \times_2 \boldsymbol{B}_i^{(2)} \times_3 \boldsymbol{B}_i^{(3)}$ follows tensor normal distribution

$$\widetilde{\boldsymbol{\mathcal{R}}}_i \backsim \boldsymbol{N}_{J,K,L}\left(\boldsymbol{\mathcal{O}}; \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T}, \boldsymbol{B}_i^{(2)}\boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T}, \boldsymbol{B}_i^{(3)}\boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T}\right)$$

Since the random effects core tensor and residual errors tensor are independent of each other, $\boldsymbol{\mathcal{E}}_i \backsim \boldsymbol{N}_{J,K,L}(\boldsymbol{\mathcal{O}}; \boldsymbol{\Sigma}_{\varepsilon}, \boldsymbol{\Psi}_{\varepsilon}, \boldsymbol{\Omega}_{\varepsilon})$, the three dimensional joint tensor $\begin{bmatrix} \widetilde{\boldsymbol{\mathcal{R}}}_i & \boldsymbol{\mathcal{O}} \\ \boldsymbol{\mathcal{O}} & \boldsymbol{\mathcal{E}}_i \end{bmatrix}$ satisfies

$$\begin{bmatrix} \widetilde{\boldsymbol{\mathcal{R}}}_i & \boldsymbol{\mathcal{O}} \\ \boldsymbol{\mathcal{O}} & \boldsymbol{\mathcal{E}}_i \end{bmatrix} \backsim \boldsymbol{N}_{2J,2K,2L}\left(\boldsymbol{\mathcal{O}}; \begin{bmatrix} \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\varepsilon} \end{bmatrix}, \begin{bmatrix} \boldsymbol{B}_i^{(2)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi}_{\varepsilon} \end{bmatrix}, \begin{bmatrix} \boldsymbol{B}_i^{(3)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}_{\varepsilon} \end{bmatrix}\right)$$

Based on the Theorem 3.1 (Ohlson et al. 2011), we define $\mathcal{A} = [\boldsymbol{I}_J \quad \boldsymbol{I}_J] \otimes [\boldsymbol{I}_K \quad \boldsymbol{I}_K] \otimes [\boldsymbol{I}_L \quad \boldsymbol{I}_L] \in \boldsymbol{\mathcal{T}}_{\otimes}^{[J,K,L],[2J,2K,2L]}$, then

$$\mathcal{A}\begin{bmatrix} \widetilde{\boldsymbol{\mathcal{R}}}_i & \boldsymbol{\mathcal{O}} \\ \boldsymbol{\mathcal{O}} & \boldsymbol{\mathcal{E}}_i \end{bmatrix} = \widetilde{\boldsymbol{\mathcal{R}}}_i + \boldsymbol{\mathcal{E}}_i$$
$$\backsim \boldsymbol{N}_{J,K,L}\left(\boldsymbol{\mathcal{O}}; \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T} + \boldsymbol{\Sigma}_{\varepsilon}, \boldsymbol{B}_i^{(2)}\boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T} + \boldsymbol{\Psi}_{\varepsilon}, \boldsymbol{B}_i^{(3)}\boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T} + \boldsymbol{\Omega}_{\varepsilon}\right)$$

Thus

$$\boldsymbol{\mathcal{Y}}_i \backsim \boldsymbol{N}_{J,K,L}\left(\left[\!\left[\boldsymbol{\mathcal{F}}; \boldsymbol{A}_i^{(1)}, \boldsymbol{A}_i^{(2)}, \boldsymbol{A}_i^{(3)}\right]\!\right]; \boldsymbol{B}_i^{(1)}\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{B}_i^{(1)^T} + \boldsymbol{\Sigma}_{\varepsilon}, \boldsymbol{B}_i^{(2)}\boldsymbol{\Psi}_{\mathrm{r}}\boldsymbol{B}_i^{(2)^T} + \boldsymbol{\Psi}_{\varepsilon}, \boldsymbol{B}_i^{(3)}\boldsymbol{\Omega}_{\mathrm{r}}\boldsymbol{B}_i^{(3)^T} + \boldsymbol{\Omega}_{\varepsilon}\right).$$

A.5 Proof of the Proposition 5

Proof:

The likelihood function for Equation (26) is shown as following:

$$L_1 = (2\pi)^{-\frac{JKL}{2}} \cdot |\boldsymbol{\Omega}_i \otimes \boldsymbol{\Psi}_i|^{-\frac{J}{2}} \cdot |\boldsymbol{\Sigma}_i|^{-\frac{KL}{2}} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)}\right)^T\right)^T (\boldsymbol{\Omega}_i \otimes \boldsymbol{\Psi}_i)^{-1}\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)}\right)^T\right)\right]\right) = (2\pi)^{-\frac{JKL}{2}} \cdot |\boldsymbol{\Omega}_i|^{-\frac{JK}{2}} \cdot |\boldsymbol{\Psi}_i|^{-\frac{JL}{2}} \cdot |\boldsymbol{\Sigma}_i|^{-\frac{KL}{2}} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)}\right)^T\right)^T (\boldsymbol{\Omega}_i \otimes \boldsymbol{\Psi}_i)^{-1}\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)}\right)^T\right)\right]\right).$$

Similarly, we can get the likelihood function for Equations (27) and (28):

$$L_2 = (2\pi)^{-\frac{JKL}{2}} \cdot |\boldsymbol{\Omega}_i|^{-\frac{JK}{2}} \cdot |\boldsymbol{\Psi}_i|^{-\frac{JL}{2}} \cdot |\boldsymbol{\Sigma}_i|^{-\frac{KL}{2}} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Psi}_i^{-1}\left(\boldsymbol{Y}_{i(2)} - \boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(1)}\right)^T\right)^T (\boldsymbol{\Omega}_i \otimes \boldsymbol{\Sigma}_i)^{-1}\left(\boldsymbol{Y}_{i(2)} - \boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(1)}\right)^T\right)\right]\right);$$

$$L_3 = (2\pi)^{-\frac{JKL}{2}} \cdot |\boldsymbol{\Omega}_i|^{-\frac{JK}{2}} \cdot |\boldsymbol{\Psi}_i|^{-\frac{JL}{2}} \cdot |\boldsymbol{\Sigma}_i|^{-\frac{KL}{2}} \cdot \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}_i^{-1}\left(\boldsymbol{Y}_{i(3)} - \boldsymbol{A}_i^{(3)}\boldsymbol{F}_{(3)}\left(\boldsymbol{A}_i^{(2)} \otimes \boldsymbol{A}_i^{(1)}\right)^T\right)^T (\boldsymbol{\Psi}_i \otimes \boldsymbol{\Sigma}_i)^{-1}\left(\boldsymbol{Y}_{i(3)} - \boldsymbol{A}_i^{(3)}\boldsymbol{F}_{(3)}\left(\boldsymbol{A}_i^{(2)} \otimes \boldsymbol{A}_i^{(1)}\right)^T\right)\right]\right).$$

To prove the log-likelihood functions of Equations (26-28) are the same, we need to show the parts within th[·] are the same.

Considering the commutation matrix $\boldsymbol{K}_{L,JK}$ and $\boldsymbol{K}_{R_1,P_1Q_1}$, we have

$$\mathrm{vec}\left(\boldsymbol{Y}_{i(1)}\right) = \boldsymbol{K}_{L,JK}\mathrm{vec}\left(\boldsymbol{Y}_{i(3)}\right)$$

$$\mathrm{vec}\left(\boldsymbol{F}_{(1)}\right) = \boldsymbol{K}_{R_1,P_1Q_1}\mathrm{vec}\left(\boldsymbol{F}_{(3)}\right)$$

$$\mathrm{vec}\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)}\right)^T\right) = \boldsymbol{K}_{K,JL}\mathrm{vec}\left(\boldsymbol{Y}_{i(2)} - \boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(1)}\right)^T\right)$$

$$\mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{Y}_{i(1)}-\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right)^T(\boldsymbol{\Omega}_i\otimes\boldsymbol{\Psi}_i)^{-1}\left(\boldsymbol{Y}_{i(1)}-\right.\right.$$

$$\left.\left.\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right)\right]=\mathrm{vec}^T\left[\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{Y}_{i(1)}-\right.\right.\right.$$

$$\left.\left.\left.\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right)^T(\boldsymbol{\Omega}_i\otimes\boldsymbol{\Psi}_i)^{-1}\right)^T\right]\mathrm{vec}\left[\boldsymbol{Y}_{i(1)}-\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right]=$$

$$\mathrm{vec}^T\left[\boldsymbol{Y}_{i(1)}-\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right]\cdot(\boldsymbol{\Omega}_i^{-1}\otimes\boldsymbol{\Psi}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})^T\cdot\mathrm{vec}\left[\boldsymbol{Y}_{i(1)}-\right.$$

$$\left.\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right]=\mathrm{vec}^T\left[\boldsymbol{Y}_{i(2)}-\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right]\boldsymbol{K}_{K,JL}^T\cdot$$

$$(\boldsymbol{\Omega}_i^{-1}\otimes\boldsymbol{\Psi}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})^T\cdot\boldsymbol{K}_{K,JL}\mathrm{vec}\left[\boldsymbol{Y}_{i(2)}-\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right]=\mathrm{vec}^T\left[\boldsymbol{Y}_{i(2)}-\right.$$

$$\left.\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right]\cdot(\boldsymbol{\Omega}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Psi}_i^{-1})^T\cdot\mathrm{vec}\left[\boldsymbol{Y}_{i(2)}-\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right]=$$

$$\mathrm{vec}^T\left[\left(\boldsymbol{\Psi}_i^{-1}\left(\boldsymbol{Y}_{i(2)}-\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right)^T(\boldsymbol{\Omega}_i\otimes\boldsymbol{\Sigma}_i)^{-1}\right)^T\right]\mathrm{vec}\left[\boldsymbol{Y}_{i(2)}-\right.$$

$$\left.\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right]=\mathrm{tr}\left[\boldsymbol{\Psi}_i^{-1}\left(\boldsymbol{Y}_{i(2)}-\right.\right.$$

$$\left.\left.\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right)^T(\boldsymbol{\Omega}_i\otimes\boldsymbol{\Sigma}_i)^{-1}\left(\boldsymbol{Y}_{i(2)}-\boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right)\right].$$

Thus, $L_1=L_2$. Similarly, we can prove $L_3=L_2$.

According to the result above, we can get the log-likelihood function as

$$\boldsymbol{l}_i=-\frac{JKL}{2}\log 2\pi-\frac{JK}{2}\log|\boldsymbol{\Omega}_i|-\frac{JL}{2}\log|\boldsymbol{\Psi}_i|-\frac{KL}{2}\log|\boldsymbol{\Sigma}_i|-\frac{1}{2}\left(\mathrm{vec}\left(\boldsymbol{Y}_{i(1)}-\right.\right.$$

$$\left.\left.\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right)\right)^T(\boldsymbol{\Omega}_i^{-1}\otimes\boldsymbol{\Psi}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathrm{vec}\left(\boldsymbol{Y}_{i(1)}-\boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right).$$

A.6 Proof of the Proposition 6

Proof:

Given the response tensors $\boldsymbol{\mathcal{Y}}_i$, the basis $\boldsymbol{A}_i^{(1)}$, $\boldsymbol{A}_i^{(2)}$, $\boldsymbol{A}_i^{(3)}$ with $i = 1,\cdots,N$, the log-likelihood function for all the samples is

$$\boldsymbol{l} = \sum_{i=1}^{N} \boldsymbol{l}_i = \sum_{i=1}^{N} \left\{ -\frac{JKL}{2}\log 2\pi - \frac{JK}{2}\log|\boldsymbol{\Omega}_i| - \frac{JL}{2}\log|\boldsymbol{\Psi}_i| - \frac{KL}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\left(\mathrm{vec}\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)} \boldsymbol{F}_{(1)} \left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)} \right)^T \right) \right)^T (\boldsymbol{\Omega}_i^{-1} \otimes \boldsymbol{\Psi}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1}) \mathrm{vec}\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)} \boldsymbol{F}_{(1)} \left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)} \right)^T \right) \right\}.$$

We take the first derivative of the log-likelihood function with respect to $\mathrm{vec}(\widehat{\boldsymbol{\mathcal{F}}})$ is

$$\frac{\boldsymbol{dl}}{\boldsymbol{d}\mathrm{vec}(\widehat{\boldsymbol{\mathcal{F}}})} = \sum_{i=1}^{N} \mathrm{vec}\left[\boldsymbol{\mathcal{Y}}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}} \right] \cdot (\boldsymbol{\Omega}_i^{-1} \otimes \boldsymbol{\Psi}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1}) \left(\boldsymbol{A}_i^{(3)} \otimes \boldsymbol{A}_i^{(2)} \otimes \boldsymbol{A}_i^{(1)} \right).$$

Let the first derivative above equals to zero, we can get the maximum likelihood estimator of $\mathrm{vec}(\boldsymbol{\mathcal{F}})$ is

$$\mathrm{vec}(\widehat{\boldsymbol{\mathcal{F}}}) = \left(\sum_{i=1}^{N} \left(\boldsymbol{A}_i^{(3)^T} \boldsymbol{\Omega}_i^{-1} \boldsymbol{A}_i^{(3)} \right) \otimes \left(\boldsymbol{A}_i^{(2)^T} \boldsymbol{\Psi}_i^{-1} \boldsymbol{A}_i^{(2)} \right) \otimes \left(\boldsymbol{A}_i^{(1)^T} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{A}_i^{(1)} \right) \right)^{-1} \cdot \left(\sum_{i=1}^{N} \left(\boldsymbol{A}_i^{(3)^T} \boldsymbol{\Omega}_i^{-1} \right) \otimes \left(\boldsymbol{A}_i^{(2)^T} \boldsymbol{\Psi}_i^{-1} \right) \otimes \left(\boldsymbol{A}_i^{(1)^T} \boldsymbol{\Sigma}_i^{-1} \right) \cdot \mathrm{vec}(\boldsymbol{\mathcal{Y}}_i) \right)$$

Moreover, we can show that the estimator $\mathrm{vec}(\widehat{\boldsymbol{\mathcal{F}}})$ given in Equation (30) is uniquely determined regardless of the parametrization of the covariance matrices. Because when $\boldsymbol{\Psi}_i, i = 1,\cdots,N$ is replaced with $\mathrm{m}\boldsymbol{\Psi}_i$ and $\mathrm{m} \in \mathbb{R}^+$, the expression $\mathrm{vec}(\widehat{\boldsymbol{\mathcal{F}}})$ above still satisfies.

Assume that $\boldsymbol{B}_i^{(1)}$, $\boldsymbol{B}_i^{(2)}$, $\boldsymbol{B}_i^{(3)}$ are constant for all $i = 1,\cdots,N$, and define that $\boldsymbol{B}_i^{(1)} = \boldsymbol{B}^{(1)}$, $\boldsymbol{B}_i^{(2)} = \boldsymbol{B}^{(2)}$, $\boldsymbol{B}_i^{(3)} = \boldsymbol{B}^{(3)}$ for $i = 1,\cdots,N$, We take the first derivatives of the log-likelihood functions with respect to $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Omega}_i$ are

$$\frac{\boldsymbol{dl}}{\boldsymbol{d\Sigma}_i} = \frac{KLN}{2}\boldsymbol{\Sigma_i^{-1}} - \boldsymbol{\Sigma_i^{-1}}\left\{\sum_{i=1}^{N}\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right)(\boldsymbol{\Omega}_i^{-1}\otimes\boldsymbol{\Psi}_i^{-1})\left(\boldsymbol{Y}_{i(1)} - \boldsymbol{A}_i^{(1)}\boldsymbol{F}_{(1)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(2)}\right)^T\right)^T\right\}\boldsymbol{\Sigma_i^{-1}}.$$

$$\frac{\boldsymbol{dl}}{\boldsymbol{d\Psi}_i} = \frac{JLN}{2}\boldsymbol{\Psi_i^{-1}} - \boldsymbol{\Psi_i^{-1}}\left\{\sum_{i=1}^{N}\left(\boldsymbol{Y}_{i(2)} - \boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right)(\boldsymbol{\Omega}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\left(\boldsymbol{Y}_{i(2)} - \boldsymbol{A}_i^{(2)}\boldsymbol{F}_{(2)}\left(\boldsymbol{A}_i^{(3)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right)^T\right\}\boldsymbol{\Psi_i^{-1}}.$$

$$\frac{\boldsymbol{dl}}{\boldsymbol{d\Omega}_i} = \frac{KJN}{2}\boldsymbol{\Omega_i^{-1}} - \boldsymbol{\Omega_i^{-1}}\left\{\sum_{i=1}^{N}\left(\boldsymbol{Y}_{i(3)} - \boldsymbol{A}_i^{(3)}\boldsymbol{F}_{(3)}\left(\boldsymbol{A}_i^{(2)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right)(\boldsymbol{\Psi}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\left(\boldsymbol{Y}_{i(3)} - \boldsymbol{A}_i^{(3)}\boldsymbol{F}_{(3)}\left(\boldsymbol{A}_i^{(2)}\otimes\boldsymbol{A}_i^{(1)}\right)^T\right)^T\right\}\boldsymbol{\Omega_i^{-1}}.$$

Letting the first derivatives of the log-likelihood functions with respect to $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Omega}_i$ be zeros, we can get the maximum likelihood estimators of $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Omega}_i$ are

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{KLN}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(1)}\cdot\left(\widehat{\boldsymbol{\Omega}}_i^{-1}\otimes\widehat{\boldsymbol{\Psi}}_i^{-1}\right)\cdot\left(\boldsymbol{y}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(1)}^T;$$

$$\widehat{\boldsymbol{\Psi}}_i = \frac{1}{JLN}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(2)}\cdot\left(\widehat{\boldsymbol{\Omega}}_i^{-1}\otimes\widehat{\boldsymbol{\Sigma}}_i^{-1}\right)\cdot\left(\boldsymbol{y}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(2)}^T;$$

$$\widehat{\boldsymbol{\Omega}}_i = \frac{1}{JKN}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(3)}\cdot\left(\widehat{\boldsymbol{\Psi}}_i^{-1}\otimes\widehat{\boldsymbol{\Sigma}}_i^{-1}\right)\cdot\left(\boldsymbol{y}_i - \widehat{\widehat{\boldsymbol{\mathcal{F}}}}\right)_{(3)}^T.$$

Straightforwardly, if both $\boldsymbol{A}_i^{(1)}$, $\boldsymbol{A}_i^{(2)}$, $\boldsymbol{A}_i^{(3)}$ and $\boldsymbol{B}_i^{(1)}$, $\boldsymbol{B}_i^{(2)}$, $\boldsymbol{B}_i^{(3)}$ are constant for all $i = 1,\cdots,N$, the maximum likelihood estimators of $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Psi}_i$, $\boldsymbol{\Omega}_i$ are

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{KLN}\sum_{i=1}^{N}(\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(1)} \cdot \left(\widehat{\boldsymbol{\Omega}}_i^{-1} \otimes \widehat{\boldsymbol{\Psi}}_i^{-1}\right) \cdot (\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(1)}^{T};$$

$$\widehat{\boldsymbol{\Psi}}_i = \frac{1}{JLN}\sum_{i=1}^{N}(\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(2)} \cdot \left(\widehat{\boldsymbol{\Omega}}_i^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_i^{-1}\right) \cdot (\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(2)}^{T};$$

$$\widehat{\boldsymbol{\Omega}}_i = \frac{1}{JKN}\sum_{i=1}^{N}(\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(3)} \cdot \left(\widehat{\boldsymbol{\Psi}}_i^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_i^{-1}\right) \cdot (\boldsymbol{\mathcal{Y}}_i - \bar{\boldsymbol{\mathcal{Y}}})_{(3)}^{T}.$$

REFERENCES

Abu‐Absi, N. R., Kenty, B. M., Cuellar, M. E., Borys, M. C., Sakhamuri, S., Strachan, D. J., . . . Li, Z. J. (2011). Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in‐line Raman spectroscopy probe. *Biotechnology and bioengineering, 108*(5), 1215-1221.

Alahbabi, M. N., Cho, Y. T., & Newson, T. P. (2006). Long-range distributed temperature and strain optical fibre sensor based on the coherent detection of spontaneous Brillouin scattering with in-line Raman amplification. *Measurement Science and Technology, 17*(5), 1082.

Bader, B. W., Kolda, T.G., and others. MATLAB Tensor Toolbox Version 2.6, Available online, (2015). URL: http://www.sandia.gov/~tgkolda/TensorToolbox/.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*: Cambridge university press.

Candes, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta numerica, 15*, 257-325.

Burg, J.P., Luenberger, D.G. and Wenger, D.L. (1982). Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9), 963-974.

Chang, S. I., & Yadama, S. (2010). Statistical process control for monitoring non-linear profiles using wavelet filtering and B-spline approximation. *International Journal of Production Research, 48*(4), 1049-1068.

Chen, T., Morris, J., & Martin, E. (2007). Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems, 87*(1), 59-71.

Cheng, Q., Bao, J., Park, J., Liang, Z., Zhang, C., & Wang, B. (2009). high mechanical performance composite conductor: multi - walled carbon nanotube sheet/bismaleimide nanocomposites. *Advanced Functional Materials, 19*(20), 3219-3225.

Cheng, Q., Wang, B., Zhang, C., & Liang, Z. (2010). Functionalized Carbon - Nanotube Sheet/Bismaleimide Nanocomposites: Mechanical and Electrical Performance Beyond Carbon - Fiber Composites. *Small, 6*(6), 763-767.

Chicken, E., Pignatiello Jr, J. J., & Simpson, J. R. (2009). Statistical process monitoring of nonlinear profiles using wavelets. *Journal of Quality Technology, 41*(2), 198.

Clegg, S. M., Sklute, E., Dyar, M. D., Barefield, J. E., & Wiens, R. C. (2009). Multivariate analysis of remote laser-induced breakdown spectroscopy spectra using partial least squares, principal component analysis, and related techniques. *Spectrochimica Acta Part B: Atomic Spectroscopy, 64*(1), 79-88.

Daubechies, I. (1992). *Ten lectures on wavelets*: SIAM.

De Beer, T., Baeyens, W., Ouyang, J., Vervaet, C., & Remon, J. P. (2006). Raman spectroscopy as a process analytical technology tool for the understanding and the quantitative in-line monitoring of the homogenization process of a pharmaceutical suspension. *Analyst, 131*(10), 1137-1144.

Demidenko, E. (2013). Mixed models: theory and applications with R. John Wiley & Sons.

Ding, Y., Zeng, L., & Zhou, S. (2006). Phase I analysis for monitoring nonlinear profiles in manufacturing processes. *Journal of Quality Technology, 38*(3), 199.

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association, 90*(432), 1200-1224.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2), 105-123.

Ehrentreich, F. (2002). Wavelet transform applications in analytical chemistry. *Analytical and Bioanalytical Chemistry, 372*(1), 115-121.

Eld′en, L. and Savas, B. (2009). A Newton–Grassmann method for computing the best multilinear rank-(r1, r2, r3) approximation of a tensor, *SIAM Journal on Matrix Analysis and Application*, 31(2), 248–271.

Fevotte, G. (2007). In situ Raman spectroscopy for in-line control of pharmaceutical crystallization and solids elaboration processes: a review. *Chemical Engineering Research and Design, 85*(7), 906-920.

Gałecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R: A step-by-step approach*: Springer Science & Business Media.

Gommans, H., Alldredge, J., Tashiro, H., Park, J., Magnuson, J., & Rinzler, A. (2000). Fibers of aligned single-walled carbon nanotubes: Polarized Raman spectroscopy. *Journal of Applied Physics, 88*(5), 2509-2514.

Jack, D., Yeh, C., Liang, Z., Li, S., Park, J., & Fielding, J. (2010). Electrical conductivity modeling and experimental study of densely packed SWCNT networks. *Nanotechnology, 21*(19), 195703.

Jensen, W. A., & Birch, J. B. (2009). Profile monitoring via nonlinear mixed models. *Journal of Quality Technology, 41*(1), 18.

Jin, J., & Shi, J. (1999). Feature-preserving data compression of stamping tonnage information using wavelets. *Technometrics, 41*(4), 327-339.

Jin, J., & Shi, J. (2001). Automatic feature extraction of waveform signals for in-process diagnostic performance improvement. *Journal of Intelligent Manufacturing, 12*(3), 257-268.

Johnstone, I. M., & Lu, A. Y. (2012). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*.

Johnstone, I. M., & Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59*(2), 319-351.

Joseph, V. R., & Hung, Y. (2008). Orthogonal-maximin Latin hypercube designs. *Statistica Sinica, 18*(1), 171.

Kang, L., & Albin, S. (2000). On-line monitoring when the process yields a linear. *Journal of Quality Technology, 32*(4), 418-426.

Kim, K., Mahmoud, M. A., & Woodall, W. H. (2003). On the monitoring of linear profiles. *Journal of Quality Technology, 35*(3), 317.

Kolda, T.G. and Bader, B.W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3), 455-500.

Liang, Z., Wang, B., Zhang, C., Ugarte, J.T., Lin, C.Y. and Thagard, J., Florida State University Research Foundation Inc, (2008). Method for continuous fabrication of carbon nanotube networks or membrane materials. U.S. Patent 7,459,121.

Lippert C., Listgarten J., Liu Y., Kadie C.M., Davidson R.I., Heckerman D. (2011). FaST linear mixed models for genome-wide association studies, *Nature Methods*, 8(10), 833-837.

Liu, T., & Kumar, S. (2003). Quantitative characterization of SWNT orientation by polarized Raman spectroscopy. *Chemical Physics Letters, 378*(3), 257-262.

Lu, N. and Zimmerman, D. (2004). On likelihood-based inference for a separable covariance matrix. Statistics and Actuarial Science Dept., Univ. of Iowa, Iowa City, IA, Tech. Rep, 337.

Mahmoud, M. A., Parker, P. A., Woodall, W. H., & Hawkins, D. M. (2007). A change point method for linear profile data. *Quality and Reliability Engineering International, 23*(2), 247-268.

Manceur, A.M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239, 37-49.

McCreery, R. L. (2005). *Raman spectroscopy for chemical analysis* (Vol. 225): John Wiley & Sons.

Mosesova, S. A., Chipman, H. A., MacKay, R. J., & Steiner, S. H. (2006). *Profile monitoring using mixed-effects models*. Retrieved from

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (Vol. 4): Irwin Chicago.

Nzabanita, J., von Rosen, D. and Singull, M. (2015). Maximum likelihood estimation in the tensor normal model with a structured mean. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A808970&dswid=-7688.

Ohlson, M., Ahmad, M.R. and von Rosen, D. (2011). More on the Kronecker structured covariance matrix. http://www.divaportal.org/smash/record.jsf?pid=diva2%3A391051&dswid=4352.

Ono, T., Ter Horst, J., & Jansens, P. (2004). Quantitative measurement of the polymorphic transformation of L-glutamic acid using in-situ Raman spectroscopy. *Crystal Growth & Design, 4*(3), 465-469.

Parikh, N., & Boyd, S. P. (2014). Proximal Algorithms. *Foundations and Trends in optimization, 1*(3), 127-239.

Park, J. G., Li, S., Liang, R., Zhang, C., & Wang, B. (2008). Structural changes and Raman analysis of single-walled carbon nanotube buckypaper after high current density induced burning. *Carbon, 46*(9), 1175-1183.

Park, J.G., Smithyman, J., Lin, C.Y., Cooke, A., Kismarahardja, A.W., Li, S., Liang, R., Brooks, J.S., Zhang, C. and Wang, B. (2009). Effects of surfactants and alignment on the physical properties of single-walled carbon nanotube buckypaper. *Journal of Applied Physics*, 106(10), 104310.

Paynabar, K., & Jin, J. (2011). Characterization of non-linear profiles variations using mixed-effect models and wavelets. *IIE Transactions, 43*(4), 275-290.

Pelletier, M. (2003). Quantitative analysis using Raman spectrometry. *Applied Spectroscopy, 57*(1), 20A-42A.

Ramos, P. M., & Ruisánchez, I. (2005). Noise and background removal in Raman spectra of ancient pigments using wavelet transform. *Journal of Raman Spectroscopy, 36*(9), 848-856.

Raravikar, N. R., Keblinski, P., Rao, A. M., Dresselhaus, M. S., Schadler, L. S., & Ajayan, P. M. (2002). Temperature dependence of radial breathing mode Raman frequency of single-walled carbon nanotubes. *Physical Review B, 66*(23), 235424.

Roś, B., Bijma, F., de Munck, J.C. and de Gunst, M.C. (2016). Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure. *Journal of Multivariate Analysis*, 143, 345-361.

Sakata, T. (2016). *Applied matrix and tensor variate data analysis*. Springer.

Salzer, R. and Siesler, H.W. (2009). *Infrared and Raman spectroscopic imaging*. John Wiley & Sons.

Shiau, J.-J. H., Huang, H.-L., Lin, S.-H., & Tsai, M.-Y. (2009). Monitoring nonlinear profiles with random effects by nonparametric regression. *Communications in Statistics—Theory and Methods, 38*(10), 1664-1679.

Snyder, D. L., White, R. L., & Hammoud, A. M. (1993). Image recovery from data acquired with a charge-coupled-device camera. *JOSA A, 10*(5), 1014-1023.

Torrence, C., & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological society, 79*(1), 61-78.

Tseng, P. (2008). On Accelerated Proximal Gradient Methods for Convex-Concave Optimization. *submitted to SIAM Journal on Optimization*.

Wahba, G. (1990). *Spline models for observational data*: SIAM.

Wang, S., Haldane, D., Liang, R., Smithyman, J., Zhang, C. and Wang, B. (2012). Nanoscale infiltration behaviour and through-thickness permeability of carbon nanotube buckypapers. *Nanotechnology*, 24(1), 015704.

Werner, K., Jansson, M. and Stoica, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2), 478-491.

Yan, H., Paynabar, K., & Shi, J. (2017). Anomaly Detection in Images With Smooth Background via Smooth-Sparse Decomposition, *Technometrics*, 59:1, 102-114.

Yue, X., Wang, K., Yan, H., Park, J.G., Liang, Z., Zhang, C., Wang, B., Shi, J. (2017). Generalized Wavelet Shrinkage of In-line Raman Spectroscopy for Quality Monitoring of Continuous Nanomanufacturing for Buckypaper. *IEEE Transactions on Automation Science and Engineering*, 14(1), 196-207.

Yue, X., Yan, H., Park, J.G., Liang, Z., Shi, J. (2018). A Wavelet-based Penalized Mixed Effects Decomposition for Multichannel Profile Monitoring Based on In-line Raman Spectroscopy, *IEEE Transactions on Automation Science and Engineering*, (in press).

Zhou, S., & Jin, J. (2005). Automatic feature selection for unsupervised clustering of cycle-based signals in manufacturing processes. *Iie Transactions, 37*(6), 569-584.

Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502), 540-552.

Zou, C., Tsung, F., & Wang, Z. (2008). Monitoring profiles based on nonparametric regression methods. *Technometrics, 50*(4), 512-526.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics, 15*(2), 265-286.